PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MARTIN SPIER

DESENVOLVIMENTO DE UM MODELO PARAMÉTRICO DE ESTIMATIVA DE ESFORÇO NO DESENVOLVIMENTO DE SOFTWARE

Porto Alegre

25 de junho de 2008

MARTIN SPIER

DESENVOLVIMENTO DE UM MODELO PARAMÉTRICO DE ESTIMATIVA DE ESFORÇO NO DESENVOLVIMENTO DE SOFTWARE

Trabalho apresentado como parte da avaliação na disciplina "Trabalho de Conclusão II" em julho de 2008, para o Curso de Bacharelado em Ciência da Computação, Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Afonso Inácio Orth

Porto Alegre

25 de junho de 2008

Dedicatória

A quem eu dedico,

A todos àqueles que acreditam que algo só é impossível até que alguem duvide e acabe provando o contrário.

Agradecimentos

A minha família, em especial meus pais, Augusto Spier e Dorothea Thobe, por terem proporcionado uma base sólida para o aprendizado durante toda minha vida, e que hoje culmina no desenvolvimento deste trabalho de conclusão de curso.

Também agradeço a todos colegas que influenciaram direta ou indiretamente este trabalho, oferecendo seu conhecimento e experiência durante todo o desenvolvimento, em especial Odorico Mendizabal, Alberto Chemale, Gustavo Becker, Elizangela Paiva e Raquel Rodrigues por todo incentivo dado.

Não menos importante, agradeço a todos amigos, que me proporcionaram momentos de descontração em meio a tantas preocupações. Sem eles, a realização deste trabalho não seria possível. Finalmente, obrigado ao Prof. Afonso Inácio Orth, orientador deste trabalho, e Prof. Hélio Radke Bittencourt por toda sabedoria e apoio prestado.

Epígrafe

"If the facts don't fit the theory, change the facts."
Albert Einstein

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas

WBS	Work Breakdown Structure
COCOMO	COnstructive COst MOdel
LOC	Lines of Code
FP	Function Points
IFPUG	Function Point Users Group
SDC	Systems Development Corporation
SLiM	Software LIfecycle Management
SEER-SEM	System Evaluation and Estimation of Resources - Software Estimating Model
SQT	Soma dos Quadrados Totais
SQE	Soma dos Quadrados Explicados
SQR	Soma dos Quadrados dos Resíduos
EP	Erro Padrão
EPA	Environmental Protection Agency

Resumo

O presente trabalho é motivado pela crescente necessidade de melhorias no processo de estimativa de projetos de software. Tais melhorias refletem positivamente em prazos de entrega, custos e qualidade do software, gerando assim, uma vantagem competitiva na crescente indústria de desenvolvimento de software.
Tendo em vista esta necessidade, este trabalho é focado no desenvolvimento de um método que venha a suprir as necessidades organizacionais atualmente existentes de sistemas que auxiliem gerentes de projetos e responsáveis por estimativas a criarem estimativas de melhor qualidade, a um pequeno custo organizacional, levando em consideração a homogeneidade entre processos de desenvolvimento de software.
O trabalho inicia com a apresentação de diversos conceitos e técnicas utilizadas atualmente pela indústria na criação de estimativas. A seguir, é apresentada a técnica de regressão linear multivariada, técnica utilizada neste trabalho na criação de modelos de estimativa. Após a apresentação de conceitos, analisa-se como tais conceitos podem ser utilizados na criação de uma ferramenta que possa suprir as necessidades descritas acima. Finalmente são realizados testes na solução proposta, utilizando dados reais de projetos a fim de analisar seu funcionamento como ferramenta de auxilio.

Palavras-chave: Estimativa; Projeto; Software; Regressão.

Abstract

The current work is motivated by growing need of improvements on the process of estimating software projects. Such improvement reflect positively on schedule, cost and software quality, thereby, creating a competitive advantage in the growing industry of software development.
Considering this need, this work is focused on the development of a method that satisfies the organizational needs currently existent of systems that help project managers and responsible for estimates to create quality estimates, with a small organizational cost, taking into consideration the homogeneity between software development processes.
The work begins with an introduction to several concepts and techniques currently used by the industry to create estimates. After this introduction, it is presented the multivariate linear regression technique, used on this work to create estimation models. After this presentation, we will analyze how such concepts can be used to create a tool that can satisfy the needs described above. Finally, the proposed solution will be tested, using real project data to analyze its behavior as a helping tool.

Key-words: Estimates; Project; Software; Regression.

1 Introdução

Neste capítulo inicial serão apresentados os problemas, tendências e, sobretudo, as justificativas para o trabalho, além dos objetivos gerais a serem atingidos.

1.1 Contexto

O crescimento acelerado na indústria de desenvolvimento de software faz com que diversas empresas busquem melhorias no processo de desenvolvimento através de padrões, normas e métodos. Estas melhorias refletem-se positivamente nos prazos de entrega, custos e qualidade do software, fazendo com que a empresa que os adote obtenha uma vantagem sobre seus concorrentes e maior chance de sucesso competitivo na indústria de desenvolvimento de software.

1.2 Motivação

Dentre as diversas melhorias de processo almejadas por empresas do setor, podemos destacar a melhoria no processo de estimativas. Estimar o tempo, esforço e data de entrega de um software é um processo delicado, o qual envolve muita sensibilidade do gerente de projetos e de todo time, exigindo um grande conhecimento do projeto e capacidade de adaptação. Este cenário, comum às empresas de desenvolvimento de software, ocasiona um aumento nas responsabilidades do gerente de projetos, que só pode contar com sua experiência profissional como ferramenta [Cam07].

As métricas e estimativas de software vêm se tornando um dos principais tópicos na Engenharia da Informação com a crescente exigência de seus consumidores pela qualidade, rapidez, comodidade e baixo custo de implantação e manutenção de software. É impossível não enxergar tais técnicas como alavanca para um produto de melhor qualidade e com custos adequados. Mas existem ainda muitas barreiras que impedem os profissionais da área de utilizarem tais técnicas, embora a literatura disponível atualmente sobre

engenharia da informação seja relativamente ampla e variada, o que nos leva ao seguinte questionamento: Por que as métricas e estimativas de software propostas para o desenvolvimento de sistemas não são fiéis à realidade e à dimensão do problema? Tais técnicas realmente acompanharam a rápida evolução do setor [lEG07]?

Dentre as inúmeras métricas e métodos de estimativa, grande parte leva em consideração somente o tempo de desenvolvimento do software, ignorando tempo de teste, escrita de documentação, treinamento de pessoal, e outras atividades inerentes ao projeto, deixando assim uma grande "brecha" no método de estimativa. Tal falha pode custar o sucesso de um projeto, caso a meta de tempo não seja cumprida.

Cada método possui suas peculiaridades, seus pontos positivos e negativos, tendo uso somente em casos específicos e possuindo diversas premissas iniciais, que muitas vezes não são facilmente obtidas. Tendo em vista estas peculiaridades, é fácil compreender que cada método tem seu uso específico, seja por um time de desenvolvimento, seja por gerentes de projeto ou outros membros. Um bom método deve levar em consideração a homogeneidade entre processos de desenvolvimento de software, adaptando-se facilmente a cada um e atendendo suas necessidades. Uma vez que as estimativas do projeto se tornem precisas o suficiente ao ponto de minimizar a preocupação com grandes erros de estimativa, elas podem produzir benefícios adicionais, tais como [McC06]:

- Melhor visibilidade de status. Uma das melhores maneiras para se monitorar o progresso de um projeto é comparar o progresso planejado com o progresso atual. Se o progresso planejado for realista, isto é, baseado em estimativas precisas, é possível monitorar o progresso conforme o plano [McC06].

- Melhor qualidade. Estimativas precisas ajudam a evitar problemas de qualidade relacionados ao *stress* gerado pela programação. Aproximadamente 40% de todos os erros de software são causados por stress. Estes erros podem ser evitados programando as atividades adequadamente e assim depositando menos *stress* sobre desenvolvedores [Gla94].

- Melhor orçamento. Estimativas precisas ajudam a criar orçamentos precisos. Uma organização que não valoriza estimativas precisas perde a habilidade de prever os custos de seus projetos [McC06].

- Identificação adiantada de riscos. Uma das oportunidades geralmente perdidas no desenvolvimento de software é a falha de corretamente interpretar o significado de uma disparidade inicial entre resultados de um projeto e estimativas do

mesmo[McC06]. Certos fins só são desejáveis se atingidos em um certo período de tempo, como por exemplo, o lançamento de um produto antes de uma concorrente.

1.3 Objetivos

Tendo como foco principal a elaboração de um ambiente capaz de gerenciar a tarefa de criação de estimativas de tempo de desenvolvimento de software, agregando qualidade às estimativas e diminuindo a pressão sobre gerentes de projetos nesta tarefa crucial para o projeto. É pretensão deste trabalho desenvolver um modelo de sistema que venha a suprir as necessidades organizacionais atualmente existentes de sistemas que auxiliem gerentes de projetos e responsáveis por estimativas a criarem estimativas de melhor qualidade, a um pequeno custo organizacional.

Tendo em vista a homogeneidade entre projetos, podemos deduzir que nem todos os projetos podem ser estimados da mesma maneira. Cada projeto possui características próprias, que devem ser levadas em consideração em cada estimativa, seja no método utilizado ou nas estimativas em si.

Tendo em vista este objetivo, torna-se necessário o uso de um modelo de estimativas adaptável baseado nas diversas características de cada projeto, tendo como foco o aumento da precisão final da estimativa e conseqüentemente o aumento de qualidade e uso.

1.3.1 Objetivos gerais

Partindo da necessidade de oferecer um serviço que possa facilitar a criação de estimativas e ao mesmo tempo torná-las mais confiáveis, objetivou-se prover um mecanismo que possibilite a criação de um modelo de estimativas adaptável aos diversos cenários atuais levando em consideração características particulares de cada projeto com fim de obter um aumento na precisão final da estimativa.

1.3.2 Objetivos específicos

A seguir são apresentamos os objetivos específicos que foram traçados para serem atingidos ao longo deste trabalho:

- Análise de conceitos-chave em estimativa de tempo de desenvolvimento de software.

- Análise de diversos métodos, técnicas e ferramentas atualmente utilizados pela indústria.
- Análise de conceitos matemáticos aplicáveis a criação de um modelo de estimativas utilizando dados históricos como base.
- Definição do processo de estimativa a ser utilizado pela solução proposta.
- Definição dos fatores que serão considerados pela solução proposta.
- Definição da técnica de modelagem a ser utilizada pela solução proposta.
- Modelagem conceitual da solução proposta.
- Definição dos dados que serão utilizados para o teste da solução proposta.
- Desenvolvimento da solução proposta.
- Teste e validação da solução desenvolvida.

1.4 Organização do texto

Este trabalho está organizado em sete capítulos. Inicialmente, neste primeiro capítulo, aborda-se uma introdução ao tema do trabalho, apresentando o contexto em que foi gerado, sua motivação e objetivos. No segundo capítulo apresentam-se diversos conceitos de estimativa. Logo após, no terceiro capítulo, são abordadas diversas técnicas de estimativa de custo de software utilizados atualmente pela indústria. No quarto capítulo trabalha-se alguns conceitos matemáticos que serão aplicados no trabalho. No capítulo cinco, analisa-se o sistema proposto e logo após, no sexto capítulo, descrevem-se os testes realizados juntamente com seus respectivos resultados. No sétimo capítulo são abordadas algumas conclusões tiradas do projeto e finalmente no oitavo capítulo, discutem-se alguns possíveis trabalhos futuros.

2 Conceitos de estimativa

Neste capítulo serão abordados alguns aspectos referentes a estimativa de tempo de desenvolvimento de software, desde sua definição básica até o detalhamento de alguns conceitos que serão utilizados neste trabalho.

2.1 O que é uma estimativa?

A estimativa de duração de uma atividade envolve avaliar a quantidade de períodos de trabalho que provavelmente serão necessários para implementar cada atividade. Uma pessoa ou grupo da equipe do projeto que estiver mais familiarizada com a natureza de uma atividade específica deve fazer ou, no mínimo, aprovar a estimativa [(PM97a].

As estimativas de duração das atividades são avaliações quantitativas da mais provável quantidade de períodos de trabalho que será requerida para se completar uma atividade [(PM97a].

A definição de estimativa pelo dicionário é: 1 - A tentativa de avaliação ou cálculo aproximado. 2 - O cálculo preliminar de custo de um projeto. 3 - Um julgamento baseado na impressão de uma pessoa; opinião [Her00].

Estritamente falando, a definição de estimativa feita pelo dicionário está correta. Uma estimativa é uma previsão de quanto tempo um projeto irá levar e quanto irá custar. Mas estimativas em projetos de software se fundem com objetivos de negócios, compromissos e controle [BAC00].

Negócios têm razões importantes para estabelecer objetivos independentes das estimativas de software. Mas o fato de um objetivo ser desejável ou mandatório não necessariamente significa que pode ser atingido [BAC00].

Enquanto um objetivo é a descrição de um fim desejável ao negócio, um compromisso é a promessa da entrega de uma funcionalidade definida com um nível de qualidade

específico em certa data. Um compromisso pode ser mais agressivo, conservador, ou ser igual a uma estimativa [BAC00].

2.1.1 Relação entre estimativas e planos

Estimativa e planejamento são tópicos relacionados, mas estimativa não é planejamento e planejamento não é estimativa.

Estimativa deve ser tratada como um processo analítico imparcial. Planejamento deve ser tratado como um processo imparcial para atingir um objetivo. Quando estimamos alguma coisa, é perigoso desejar uma resposta em particular, pois o objetivo de uma estimativa é atingir uma precisão e não atingir um resultado. Mas o objetivo do planejamento é atingir um resultado em particular, planejando os meios para atingir um fim.

Estimativas formam a fundação dos planos, mas os planos não necessitam ser exatamente iguais as estimativas, se as estimativas são dramaticamente diferentes dos planos, estes devem ser modificados para acomodar um menor risco de fracasso do projeto.

Ambos planos e estimativas são importantes para o projeto, mas deve ser levado em consideração que a combinação dos dois tende levar a estimativas e planos pobres. A presença de um alvo de planejamento muito forte pode influenciar as estimativas, e muitas vezes o alvo pode ser referenciado como estimativa.

Abaixo são citados alguns exemplos de considerações de planejamento que dependem de estimativas precisas:

- Criação de um cronograma detalhado
- Identificação do caminho crítico do projeto
- Criação de uma WBS completa
- Priorização das funcionalidades para entrega
- Quebra do projeto em iterações

Estimativas precisas apóiam um trabalho melhor em cada uma destas áreas.

2.1.2 Estimativas e Probabilidade

Esta é uma das questões centrais de estimativa de software. Estimativas de software são geralmente apresentadas como pontos únicos, como "Este projeto terá 14 semanas". Tal estimativa simples é sem sentido, pois não inclui nenhuma indicação de probabilidade.

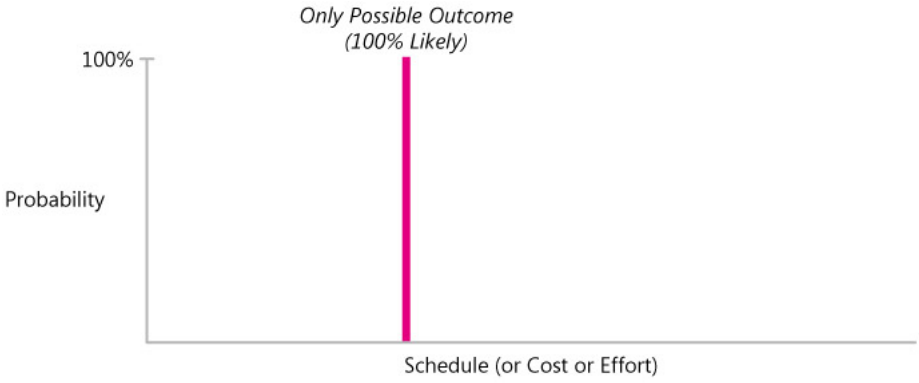


Figura 2.1: Resultado do projeto em ponto único [McC06]

Estimativas de software precisas reconhecem que projetos de software são cercados de incertezas. Coletivamente, estas diversas fontes de incerteza dizem que os resultados do projeto seguem uma distribuição de probabilidade. Como pode ser visto na Figura 2.2, a distribuição de probabilidade pode parecer como uma simples curva em sino.

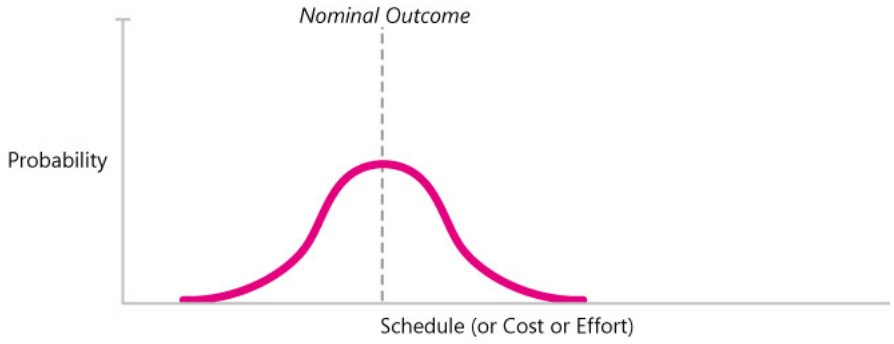


Figura 2.2: Resultado do projeto como curva em sino [McC06]

Cada ponto da curva representa a chance do projeto acabar exatamente na data estimada, ou com um custo específico. Este tipo de distribuição de probabilidade reconhece uma vasta gama de resultados. Mas a suposição de que resultados são distribuídos simétricamente ao redor de um ponto médio não é válida. Existe um limite de quão bem um projeto pode ser conduzido, o que quer dizer que a cauda esquerda da distribuição é

truncada ao invés de extendida como é feita na curva em sino. Enquanto há um limite no quão bem o projeto pode ir, não existe limite para quanto um projeto possa ser conduzido pobremente, logo a distribuição de probabilidade possui uma cauda mais longa no lado direito, como pode ser visto na Figura 2.3.

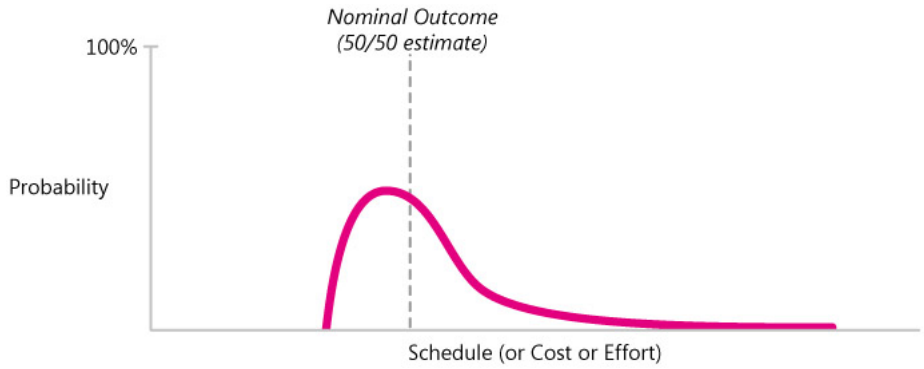


Figura 2.3: Resultado do projeto como uma distribuição mais realista [McC06]

2.1.3 Definições comuns de uma "boa" estimativa

A resposta para a questão de o que é uma estimativa ainda nos deixa com uma questão, o que é uma "boa" estimativa. Especialistas propuseram várias definições para uma "boa" estimativa. Capers Jones indicou que precisões de até 10% são possíveis, mas somente em projetos bem controlados [Jon98]. Projetos caóticos têm uma variância muito grande para atingir este nível de precisão.

Em 1986, os professores S.D. Conte, H.E. Dunsmore, e V.Y. Shen propuseram que uma abordagem de estimativa boa deve prover estimativas que estão entre 25% dos resultados atuais em 75% do tempo [CDS86]. Este padrão de avaliação é o mais comum para avaliar precisão de estimativas [Stu05].

2.1.4 Estimativas e controle do projeto

Estimativa de software é tratada normalmente como uma atividade puramente preditiva, mas a realidade é um pouco diferente. Uma vez que realizamos uma estimativa, e com base nesta estimativa, nos comprometemos em entregar funcionalidade e qualidade em uma particular data, necessitamos controlar o projeto. Atividades típicas de controle de projeto incluem remover escopo, redefinir requisitos, substituir equipe, e assim por diante.

Adicionalmente ao controle de atividades do projeto, estes projetos são normalmente afetados por eventos externos não previstos. Eventos como este acontecem durante o projeto e em muitos casos invalidam suposições que foram utilizadas para criar as estimativas em primeiro lugar. Suposições de funcionalidade mudam, suposições de alocação mudam, prioridades mudam, e assim por diante. Deste modo, fica difícil realizar uma avaliação analítica sobre se o projeto foi estimado precisamente, já que o produto que foi entregue pode não ser o mesmo que foi estimado.

Na pratica, se entregamos um projeto com aproximadamente o mesmo nível de funcionalidades, utilizando aproximadamente o mesmo nível de recursos e na mesma faixa de tempo, podemos dizer que o projeto "atingiu suas estimativas", apesar de todas as impurezas analíticas implícitas na estimativa.

2.1.5 Significado real de uma estimativa

Gerentes de projeto geralmente encontram uma discrepância entre os objetivos de negócio, seu cronograma estimado e custo. Se a diferença é pequena, o gerente pode controlar o projeto para uma conclusão com sucesso, preparando-se cautelosamente, "apertando" o cronograma, verba ou requisitos. Se a diferença é muito grande, os objetivos do projeto devem ser reconsiderados.

O objetivo primordial da estimativa de software não é prever os resultados de um projeto, e sim determinar se os objetivos do projeto são realistas o bastante para permitir que o projeto seja controlado para atingi-los.

Estimativas não precisam ser perfeitamente precisas tanto quanto precisam ser úteis. Quando temos uma combinação de estimativas precisas, bom alinhamento de objetivos e bom planejamento e controle, podemos obter resultados que são muito próximos das estimativas.

Como citado por Steve McConnel [McC06],

Uma boa estimativa é uma estimativa que provê uma visão limpa o bastante da realidade do projeto para permitir a liderança do projeto realizar boas decisões sobre como controlar o projeto para atingir seus fins.

2.2 De onde vem o erro?

O desenvolvimento de software é um processo de refinamento gradual. Se inicia com um conceito geral de um produto, a visão do software que se deseja construir, e este conceito é refinado baseado nos objetivos do projeto até que se obtém o produto final. Algumas vezes o objetivo é estimar a verba e cronograma necessários para entregar uma certa quantidade de funcionalidades. Em outros casos, o objetivo é estimar quanta funcionalidade pode ser entregue com uma quantidade de tempo determinada dentro de uma verba fixa. Projetos geralmente tem uma flexibilidade de verba, cronograma e funcionalidades a serem entregues. Esta flexibilidade dá ao projeto "caras" diferentes, diferentes combinações de custo, cronograma e conjunto de funcionalidades. Diferenças potenciais de como uma simples funcionalidade é especificada, analisada e implementada pode introduzir diferenças cumulativas que irão interferir diretamente no tempo de implementação de cada funcionalidade. Quando combinamos estas incertezas de diversas funcionalidades, podemos acabar com uma notável incerteza no projeto.

2.2.1 O cone da incerteza

Incerteza na estimativa de resultados de desenvolvimento de software vem da incerteza de como decisões necessárias ao projeto serão resolvidas. Conforme realizamos estas decisões, diminuímos o grau de incerteza do projeto. Como resultado deste processo de resolução de decisões, pesquisadores descobriram que estimativas de projeto estão sujeitas a graus de incerteza previsíveis em vários estágios do projeto. O cone da incerteza, que pode ser visto na Figura 2.4, demonstra como as estimativas tornam-se mais precisas durante as fases do projeto.

O eixo horizontal contém *milestones* comuns a projetos, como conceito inicial, definição do produto aprovada, requisitos completos, e assim por diante.

O eixo vertical contém o grau de erro que foi encontrado em estimativas criadas por avaliadores experientes em vários pontos do projeto.

Como se pode notar, estimativas criadas muito cedo no projeto são sujeitas a um grau maior de erro.

Uma questão que é levantada com freqüência é a de que caso seja disponibilizado mais tempo para realizar as estimativas, elas poderiam conter menos incerteza. Esta possibilidade, a primeira vista, parece razoável, mas infelizmente isto não é verdade. Uma

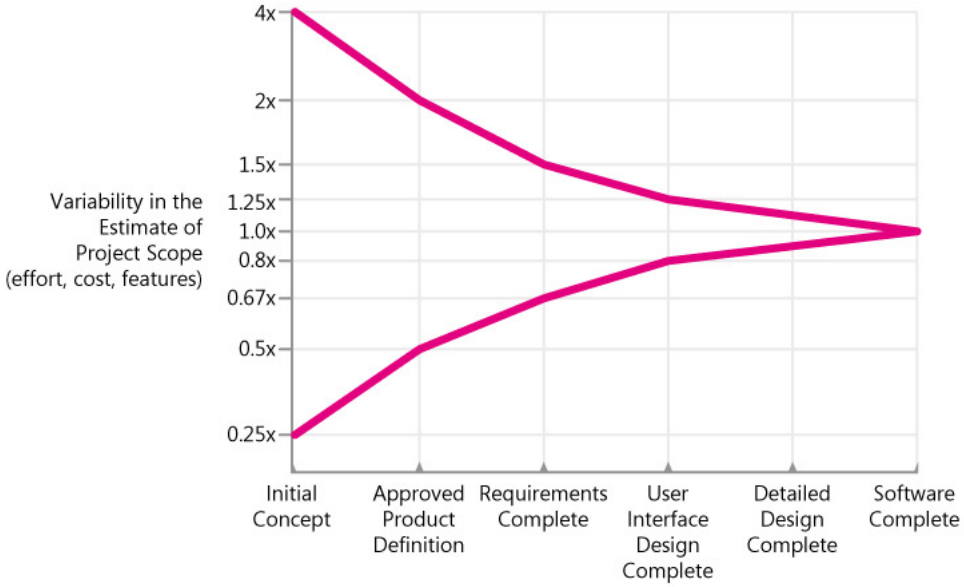


Figura 2.4: Cone da incerteza baseado em *milestones* comuns de projeto [McC06]

pesquisa realizada por Luiz Laranjeira [Lar90], sugere que a precisão da estimativa de software depende do nível de refinamento da definição do software, e não do tempo que é levado para realizar a estimativa. Quanto mais refinada a definição do software for, mais precisa será a estimativa. A razão pela qual a estimativa possui variabilidade é a de que o próprio projeto possui variabilidade, logo, a única maneira de se reduzir a variabilidade da estimativa é reduzindo a variabilidade do projeto.

Um erro comum na interpretação do cone da incerteza é a de que ele aparentemente demora tempo demais para "afinar", como se somente fosse possível obter uma estimativa refinada no final do projeto. Felizmente, esta interpretação é criada pela idéia de que os "milestones" no eixo horizontal estão igualmente espaçados, e naturalmente assumimos que o eixo horizontal é um tempo fixo. Na realidade, os "milestones" listados tendem a ser completados mais cedo no projeto. Quando o cone é desenhado novamente, com o eixo horizontal com tempo fixo, o cone da incerteza passa a ter a forma apresentada na Figura 2.5.

Um conceito mais complexo de se entender é de que o cone da incerteza representa o melhor caso de precisão que é possivel obter em diferentes pontos do projeto. O cone representa a incerteza em estimativas criadas por avaliadores experientes. É muito fácil se obter um cone "pior", mas não é possível ser mais preciso, somente ter mais sorte. Um exemplo de como o cone da incerteza pode ser "pior" é representado na Figura 2.6, que demonstra o melhor caso de estimativa, caso um projeto não seja bem controlado, ou os

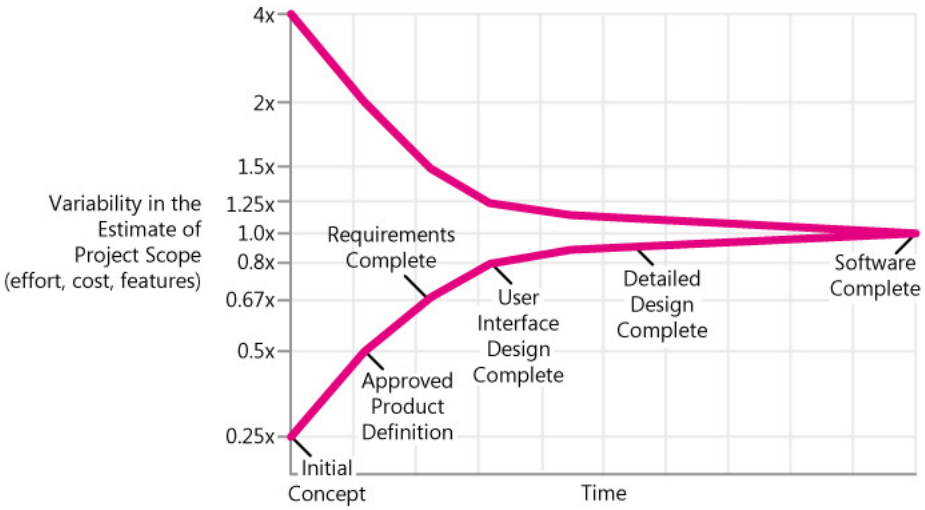


Figura 2.5: Cone da incerteza baseado em tempo fixo [McC06]

avaliadores não tenham muita prática. Neste caso, as estimativas falham em melhorar e a variabilidade não é reduzida. Ao invés de um cone, obtemos uma "nuvem" que persiste até o fim do projeto. Nova variabilidade pode ser adicionada em projetos mal controlados, como requisitos mal compreendidos, falta de envolvimento na aprovação de requisitos, análise mal feita, práticas de desenvolvimento ruins, pessoal inexperiente, entre outros.

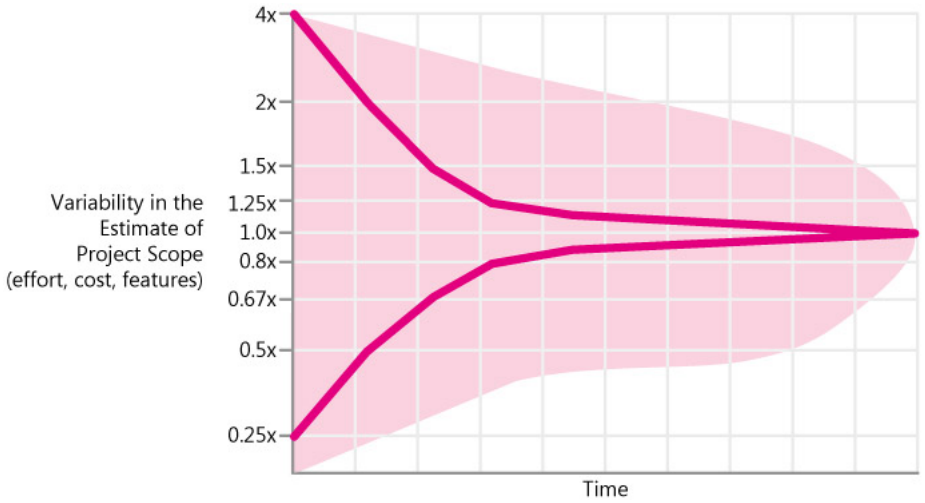


Figura 2.6: Cone da incerteza com uma "nuvem" de incerteza [McC06]

2.3 Influências

As influências de uma estimativa de tempo de desenvolvimento de software podem ser divididas de diferentes maneiras. Entender cada uma destas influência é importante

para facilitar o entendimento da dinâmica do projeto e assim melhorar a precisão das estimativas em si.

O tamanho do projeto é considerado o maior determinante de esforço, cronograma e custo de um projeto. O tipo de software a ser desenvolvido vem em segundo lugar e logo após os fatores pessoais [McC06].

2.3.1 Tamanho do projeto

O maior denominador de uma estimativa de tempo de desenvolvimento é o tamanho do projeto a ser desenvolvido, devido a sua grande variância comparado a outros fatores. Esta idéia pode parecer óbvia, mas em muitos casos, organizações violam este fato fundamental de duas maneiras.

- Custo, esforço e cronograma são estimados sem o conhecimento do tamanho do software a ser construido.
- Custo, esforço e cronograma não são ajustados quando o tamanho do software aumenta, isto é, mais escopo é adicionado ao projeto.

2.3.1.1 Deseconomias de escala

Utilizando um pensamento natural, podemos assumir que um projeto dez vezes maior que outro irá necessitar aproximadamente dez vezes mais esforço, mas na maioria dos casos, isto não é verdade. O fator básico por trás desta afirmação é de que projetos maiores necessitam mais coordenação, com um grupo maior de pessoas, o que requer mais comunicação [Jr.95]. Enquanto o número de pessoas em um projeto aumenta, o número de canais de comunicação também aumenta exponencialmente, gerando mais *overhead* para o projeto, como pode ser visto na Figura 2.7.

A consequência deste aumento exponencial nos canais de comunicação é a de que projetos também tem um crescimento exponencial em esforço. Isto é conhecido como deseconomia de escala.

A Figura 2.8 demonstra uma típica deseconomia de escala comparado ao aumento de esforço que seria associado a um crescimento linear.

Grande esforço tem sido feito na determinação exata da significância das deseconomias de escala na estimativa de software. Apesar de ser um fator significante, o tamanho do software ainda é o maior fator para criação de uma estimativa.

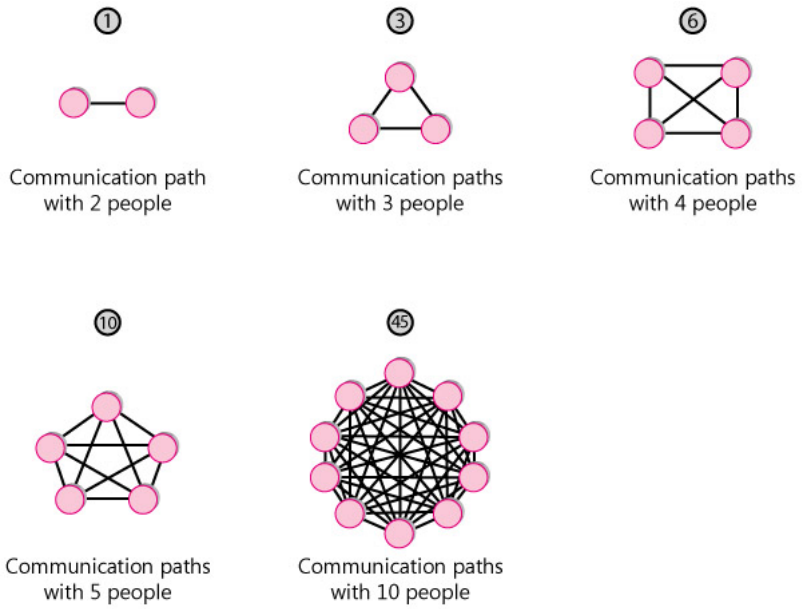


Figura 2.7: Canais de comunicação em um projeto [McC06]

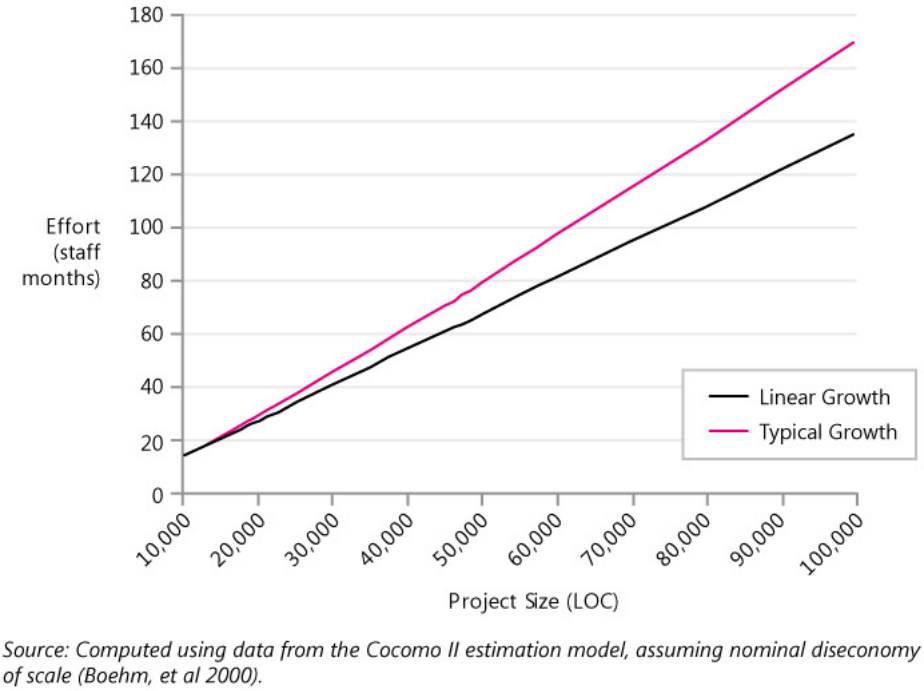


Figura 2.8: Deseconomia de escala em sistemas de negócios típicos [McC06]

2.3.2 Tipo de software a ser desenvolvido

Após o tamanho do software, o tipo de software a ser desenvolvido é o fator de maior influência nas estimativas. A Tabela 2.1 [PM92, PM97b, PM03] nos dá um exemplo de como o tipo de software a ser desenvolvido exerce influência no numero de linhas de

código por mês de trabalho.

LOC/Mês de trabalho Alto-Baixo (Nominal)			
Tipo de Software	**Projeto com 10.000 LOC**	**Projeto com 100.000 LOC**	**Projeto com 250.000 LOC**
Aviação	100-1.000 (200)	20-300 (50)	20-200 (40)
Sistemas de Negócios	800-18.000 (3.000)	200-7.000 (600)	100-5.000 (500)
Controle e Comando	200-3.000 (500)	50-600 (100)	40-500 (80)
Sistemas Embarcados	100-2.000 (300)	30-500 (70)	200-400 (60)
Sistemas de Internet	600-10.000 (1.500)	100-2.000 (300)	100-1.500 (200)
Sistemas de Intranet	1.500-18.000 (4.000)	300-7.000 (800)	200-5.000 (600)
Microcódigo	100-800 (200)	20-200 (40)	20-100 (30)
Controle de Processo	500-5.000 (1.000)	100-1.000 (300)	80-900 (200)
Sistema de Tempo Real	100-1.500 (200)	20-300 (50)	20-300 (40)
Sistemas Científicos	500-7.500 (1.000)	100-1.500 (300)	80-1.000 (200)
Sistemas de Pacotes	400-5.000 (1.000)	100-1.000 (200)	70-800 (200)
Drivers	200-5.000 (600)	50-1.000 (100)	40-800 (90)
Telecom	200-3.000 (600)	50-600 (100)	40-500 (90)

Tabela 2.1: Taxas de produtividade normais para tipos de projeto [PM92, PM97b, PM03]

Como pode ser visto na tabela, um time desenvolvendo um sistema de intranet gera código dez a vinte vezes mais rápido do que um time trabalhando em sistemas de aviação.

2.3.3 Fatores pessoais

Fatores pessoais também exercem uma influência significante nos resultados do projeto.

Dependendo da variância de cada um destes fatores, o resultado do projeto pode variar conforme é indicado na Figura 2.9, isto é, um projeto com os piores analistas de requisitos deve necessitar 42% mais esforço do que um projeto com os melhores analistas, que necessitaria 29% menos esforço, por exemplo. A magnitude destes fatores, retirados do modelo COCOMO II é confirmada por diversos estudos deste a década de sessenta [SEG68, WS74, Cur81, Mil83, BGS84, DL85, CSB+86, Car87, Boe87, BP88, VM89, Boe00].

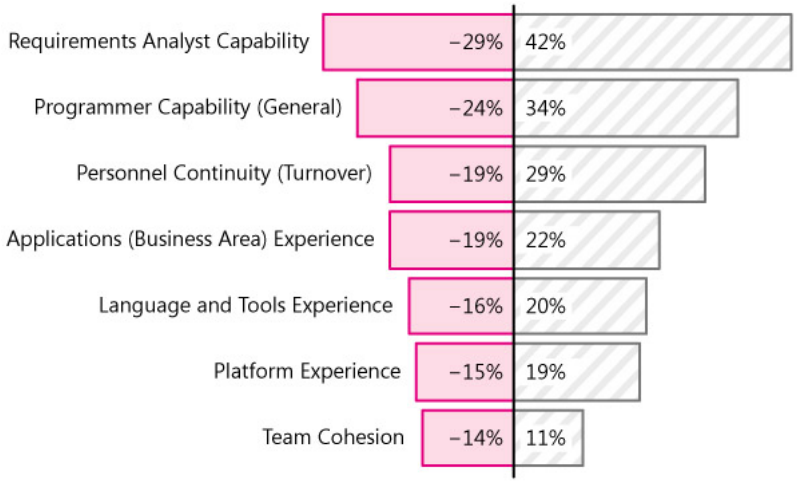


Figura 2.9: Efeito de fatores pessoais no esforço do projeto [McC06]

Uma implicação destas variações entre indivíduos é de que não se pode estimar precisamente um projeto se não temos conhecimento de quem irá executar o trabalho.

2.3.4 Linguagem de programação

A linguagem de programação que será utilizada no projeto pode afetar as estimativas em pelo menos quatro sentidos.

- A experiência que o time tem com determinada linguagem e ferramentas que serão utilizadas no projeto tem aproximadamente 40% de impacto na taxa de produtividade do projeto em geral.
- Algumas linguagens geram mais funcionalidade por linha de código do que outras. Caso não exista uma possibilidade de escolha de linguagem de programação a ser utilizada, este ponto é irrelevante, mas caso esta possibilidade exista, a utilização de linguagens que gerem mais funcionalidades por linha de código é preferencial.
- Possibilidades de ferramentas de suporte e ambientes associados com a linguagem. A escolha de linguagem pode determinar a escolha das ferramentas e ambiente.
- Desenvolvedores trabalhando em linguagens interpretadas tendem a ser mais produtivos do que aqueles trabalhando com linguagens compiladas [Jon86, Pre00].

2.3.5 Outras influências

Embora os fatores citados acima sejam os maiores determinantes de esforço de um projeto, outros fatores também podem influenciar o esforço necessário para realizar uma tarefa, entre eles, podemos citar alguns:

- Conhecimento na área de negócio da aplicação
- Tamanho da base de dados
- Desenvolvimento para reuso
- Documentação extensa
- Experiência na linguagem e ferramentas
- Desenvolvimento *multisite*
- Rotatividade de pessoal
- Experiência na plataforma
- Volatilidade da plataforma
- Complexidade do produto
- Capacidade do programador (geral)
- Confiabilidade necessária
- Capacidade de analise de requisitos
- Limitação de armazenamento
- Limitação de tempo
- Uso de ferramentas de software

2.4 Calibragem e dados históricos

Estimativas sempre envolvem algum tipo de calibragem, seja implícita ou explícita, possibilitando assim a criação de uma estimativa baseada nos dados do projeto. Estimativas podem ser calibradas utilizando dados de três tipos:

- Dados da indústria, que se referem a dados de outras organizações que desenvolvem o mesmo tipo básico de software que está sendo estimado.
- Dados históricos, que se referem a dados da organização que irá conduzir o projeto.
- Dados do projeto, que se refere a dados gerados previamente no mesmo projeto que está sendo estimado.

Dados históricos e dados do projeto são ambos muito úteis e podem apoiar a criação de estimativas altamente precisas. Do outro lado, temos os dados da indústria, que podem servir como dados temporários quando não se possui dados históricos ou dados do projeto.

2.4.1 Precisão e outros benefícios de dados históricos

A razão primordial para o uso de dados históricos na realização de estimativas é sua capacidade de gerar estimativas mais precisas. Os ítens descritos abaixo abordam algumas razões pelas quais dados históricos tendem a gerar uma estimativa mais precisa.

- **Leva em consideração influências organizacionais:** O uso de dados históricos reflete diversas influências organzacionais que afetam os resultados do projeto nas estimativas. Para projetos pequenos, capacidades individuais ditam os resultados do projeto. Conforme o tamanho de um projeto cresce, talento individual ainda conta, mas seus esforços são diluídos pelas influências organizacionais. Para projetos grandes e médios, características organizacionais começam a importar mais do que capacidades individuais [McC06].
- **Evita subjetividade e otimismo infundado:** Uma maneira pela qual subjetividade influencia negativamente as estimativas é a idéia de que um projeto novo tende a ser melhor que o anterior, influenciando assim as estimativas. Com dados históricos, se parte do princípio de que um novo projeto deve ser executado aproximadamente da mesma maneira que um anterior. Esta é uma afirmativa válida, se levar-mos em consideração a colocação de Lawrence Putnam que diz que a produtividade é um fator organizacional que não varia facilmente de projeto para projeto [PM92].
- **Reduz políticas de estimativa:** Geralmente gerentes de projetos e avaliadores têm de fazer suposições sobre a produtividade do seu time para calibrar suas estimativas, e isto pode se tornar complicado quando é preciso assumir uma posição

ruim, como uma baixa taxa de produtividade de um time. O uso de dados históricos evita esta discussão, de um time estar acima ou abaixo da média. Produtividade é o que os dados históricos nos dizem.

2.4.2 Dados a serem coletados

Diversos dados históricos podem ser coletados de projetos, desde dados estatísticos até dados subjetivos, como lições aprendidas. Dentre estes dados, podemos citar alguns que facilmente podemos utilizar na calibragem de nossas estimativas:

- Tamanho, seja em linhas de código, pontos de função ou qualquer outra medida que possa ser realizada após o termino de um projeto. Vale lembrar que esta técnica de medida deve ser utilizada em todos projetos para que possa ser comparada a outros projetos.
- Esforço, é o atual tempo que os integrantes do time necessitaram para realizar determinada tarefa. Pode ser medido em horas de trabalho.
- Tempo, é o tempo que o projeto demorou para ser desenvolvido. Pode ser medido em meses no calendário.
- Defeitos, classificados por severidade. Esta é uma medida muito útil para se determinar a qualidade do software que foi desenvolvido.

2.4.3 Utilizando dados do projeto para refinar estimativas

Como dito acima, dados históricos são úteis pois levam em consideração influências organizacionais, a mesma idéia se aplica ao uso de dados históricos dentro de um projeto específico [Gil88, Coh06]. Projetos possuem dinâmicas individuais que irão variar da dinâmica da organização que os desenvolve, deste modo, a utilização de dados históricos do projeto leva em consideração as influências únicas de cada projeto.

2.4.4 Utilizando dados da indústria para refinar estimativas

Caso dados históricos não existam, não há muita escolha a não ser o uso de dados médios da indústria, que podem se adequar ou não ao projeto sendo estimado. Nestes casos, tende a se utilizar estes dados por um período de tempo até que dados históricos são formados. Em certos casos, pode-se também utlilizar o julgamento de especialistas

na criação das estimativas, até que dados históricos são formados. Estudos indicam que a precisão encontrada em estimativas utilizando julgamento de especialistas foram mais precisas do que estimativas geradas por modelos não calibrados com dados históricos, mas modelos calibrados tendem ser tão bons quanto ou melhores do que estimativas de especialistas [Jor04].

3 Técnicas de estimativa

Diversas técnicas de estimativa foram desenvolvidas nos últimos anos, neste capítulo abordam-se algumas destas técnicas e modelos utilizados atualmente. Inicialmente serão apresentadas duas medidas de tamanho de software, e logo após, diversos técnicas de criação de estimativas.

3.1 Medidas de tamanho de software

As medidas de tamanho de software surgiram com o objetivo de estimar o esforço e o prazo associados ao desenvolvimento de sistemas [Agu03].

As medidas de tamanho de software mais comuns, como linhas de código e pontos de função tem seus próprios pontos fortes e fracos, assim como medidas customizadas definidas por organizações para uso interno [McC06].

Criar estimativas utilizando diferentes medidas de tamanho e após isto, procurar por uma convergência, tende a produzir resultados mais precisos [McC06].

3.1.1 Linhas de código

Durante bastante tempo a principal medida de tamanho de software utilizada foi a quantidade de linhas de código, ou *Lines of Code* (LOC). Há várias formas de contagem de linhas de código, algumas delas voltadas às linhas de código propriamente ditas, outras voltadas às declarações e comandos de programação contidos em uma unidade de software. A medida de linhas de código é considerada uma medida física do tamanho do software, por medir literalmente o volume de código contido no mesmo [Agu03].

A medida de linhas de código é a medida de tamanho mais utilizada para estimativa de tamanho de software.

Utilizar linhas de código pode apresentar várias vantagens, como:

- Dados sobre linhas de código de projetos antigos podem ser coletados facilmente

utilizando-se ferramentas apropriadas.

- Muitos dados históricos já existem em forma de linhas de código em várias organizações.
- Esforço por linha de código tem sido considerado aproximadamente constante através de linguagens de programação, ou parecidos o bastante para uso prático.
- Medidas via linhas de código proporcionam comparações entre projetos e estimativas de futuros projetos baseados em dados de projetos passados.
- A maioria das ferramentas comerciais de estimativa tem como base de tamanho linhas de código.

Mas por outro lado, a medida por linhas de código apresenta diversas dificuldades quando utilizada para estimar tamanho:

- Modelos simples, como linhas de código por meses de trabalho são sucetíveis a erros devido a vasta diferença entre tipos de software.
- Linhas de código não pode ser utilizado para estimar o esforço de um individuo, devido as diferenças de produtividade entre diferentes desenvolvedores..
- Projetos que necessitam mais complexidade de código do que os projetos utilizados para calibrar os fatores de produtividade podem enfraquecer a precisão da estimativa.
- A utilização de linhas de código como base para estimativa de requisitos, *design*, e outras atividades que precedem a criação de código é contra-intuitiva.
- Linhas de código são difíceis de estimar diretamente e devem ser estimadas utilizando alguma ferramenta.
- O que exatamente constitui uma linha de código deve ser definido com cuidado para evitar problemas na coleta de dados.

Apesar de seus problemas, para a maioria das organizações, a medida via linhas de código é uma boa opção para medir o tamanho de projetos passados, devido a sua simplicidade e fácil interpretação, e assim, utilizando esta base de dados para a criação de estimativas iniciais para novos projetos.

Embora a medida de linhas de código seja útil em muitos contextos, suas limitações levaram à criação de outras medidas. Essas novas medidas procuravam medir a funcionalidade disponibilizada pelo software, ao invés do tamanho físico. Por essa razão são chamadas medidas funcionais de tamanho. São úteis para produzir estimativas no início do projeto, quando pode ser muito difícil estimar a quantidade de linhas de código. A mais importante destas medidas foi apresentada por Allan Albrecht em 1979 [AG83], os pontos de função, ou Function Points (FP).

3.1.2 Pontos de função

Uma alternativa para a medida de linhas de código é a de pontos de função. Um ponto de função é uma medida sintética de tamanho de um programa que pode ser utilizada para estimar o tamanho de um projeto nas fases iniciais [Alb79]. Pontos de função são mais simples de calcular a partir de especificações de requisitos do que linhas de código, e que posteriormente podem servir de base para uma conversão. Existem diversos métodos de cálculo de pontos de função, mas o método padrão é mantido pela *International Function Point Users Group*, ou IFPUG, e pode ser encontrado no seu website [Web07].

O número de pontos de função em um programa é baseado no número e complexidade de cada um dos seguintes itens:

- **Entradas Externas:** Telas, formulários, caixas de diálogo, ou sinais de controle pelos quais um usuário final ou outro programa adiciona, deleta, ou modifica dados do programa. Elas incluem qualquer entrada que tem um formato único ou lógica de processamento única.
- **Saídas Externas:** Telas, relatórios, gráficos, ou sinais de controle que o programa gera para o uso de um usuário final ou outro programa. Elas incluem qualquer saída que tem formato diferente ou necessite uma lógica de processamento diferente de outros meios de saída.
- **Consultas Externas:** Combinações de entradas e saídas onde os resultados de uma entrada são imediatos e simples. O termo se originou da área de banco de dados e se refere a consulta direta por dados específicos. Em interfaces modernas e aplicações voltadas a web, a linha entre consultas e saídas externas é tênue, mas geralmente consultas são realizadas diretamente no banco de dados e provém uma

formatação rudimentar. Saídas externas podem processar, combinar ou sumarizar dados complexos, pondendo ser altamente formatadas.

- **Arquivos Lógicos Internos:** Maiores grupos lógicos de dados de usuários ou informações de controle que são completamente controlados pelo programa. Um arquivo lógico pode consistir de um único arquivo ou uma única tabela em um banco de dados relacional.

- **Arquivos de Interface Externos:** Arquivos controlados por outros programas com os quais o programa a ser computado interage. Isto inclui cada grande grupo lógico ou informação de controle que entra ou sai do programa.

A Tabela 3.1 [Jon91] demonstra como a contagem de entradas e saídas é convertida para pontos de função não ajustados.

Pontos de Função			
Característica do Programa	**Complexidade Baixa**	**Complexidade Média**	**Complexidade Alta**
Entradas Externas	x 3	x 4	x 6
Saídas Externas	x 4	x 5	x 7
Consultas Externas	x 3	x 4	x 6
Arquivos Lóogicos Internos	x 4	x 10	x 15
Arquivos de Interface Externos	x 5	x 7	x 10

Tabela 3.1: Multiplicadores para computação de pontos de função não ajustados [Jon91]

Entradas de complexidade baixa são multiplicadas por três, entradas de complexidade média são multiplicadas por quatro, e assim por diante. A soma destes números resulta nos pontos de função não ajustados.

Após calcular os pontos de função não ajustados, um multiplicador de influência deve ser aplicado, baseado na influência que quatorze fatores tem sobre o programa. Estes fatores incluem comunicação de dados, entrada de dados, complexidade de processamento e facilidade de instalação. A influência dos multiplicadores varia em uma faixa de 0.65 a 1.35. Quando se multiplica o total de pontos de função não ajustados pelo multiplicador de influência se obtém o total de pontos de função ajustados.

Dois estudos chegaram a uma conclusão de que pontos de função não ajustados são mais correlacionados com o tamanho final do programa do que os pontos de função ajustados [Kem87, GW91]. Alguns especialistas também recomendam a eliminação de

julgamentos de pontos de função tidos como de baixa e alta complexidade, classificando assim, todos ítens contados como de complexidade média, o que elimina outro ponto de subjetividade [Jon91].

3.1.3 Conversão entre linhas de código e pontos de função

A conversão entre linhas de código e pontos de função é simples, utilizando fatores de conversão para diversas linguagens populares. A Tabela 3.2 [Jon98, Boe00, Stu05] lista diversos fatores de conversão para linguagens populares:

Fatores de Conversão por Linguagem			
Linguagem	**Mínimo**	**Normal**	**Máximo**
Ada 83	45	80	125
Ada 95	30	50	70
C	60	128	170
C#	40	55	80
C++	40	55	140
Cobol	65	107	150
Fortran 90	45	80	125
Fortran 95	30	71	100
Java	40	55	80
Macro Assembly	130	213	300
Perl	10	20	30
Padrão de Segunda Geração	65	107	160
Smalltalk	10	20	40
SQL	7	13	15
Padrão de Terceira Geração	45	80	125
Microsoft Visual Basic	15	32	41

Tabela 3.2: Fatores de conversão para linhas de código por linguagem [Jon98, Boe00, Stu05]

Os fatores de conversão apresentados na Tabela 3.2 são exemplos de padrões da indústria. Como em vários outros fatores que podem ser estimados, a coleta de dados históricos sobre como pontos de função são transformados em linhas de código em cada organização pode ajudar a calibrar estes fatores para que seja possível estimar mais precisamente e com menor variância o tamanho do software.

3.2 Julgamento de especialistas

Julgamento de especialistas são úteis na falta de dados empíricos e quantificáveis. Elas capturam o conhecimento e experiência de participantes selecionados dentro de um domínio, provendo estimativas baseadas na síntese do conhecimento de projetos passados dos quais especialistas participaram ou tem conhecimento [BAC00].

3.2.1 Julgamento individual de especialistas

Julgamento individual de especialistas é abordagem de estimativa mais comum nos dias de hoje [Jor02]. Hihn e Habib-agahi descobriram que 83% dos projetos analisado utilizavam "analogia informal" como técnica primária de estimativa [HHA91]

Quando discutimos "julgamento de especialistas", a primeira pergunta que devemos fazer é: "Especialista em que?" Ser especialista na tecnologia ou práticas de desenvolvimento que serão empregadas não faz de uma pessoa um especialista em estimativa. Magne Jorgensen identificou que o aumento na experiência da atividade a ser estimada não leva a um aumento na precisão das estimativas das atividades em questão [Jor02].

Para a estimativa de tarefas específicas, como o tempo necessário para codificar uma funcionalidade, as pessoas que irão executar a tarefa tendem a criar as estimativas mais precisas [LP92].

Julgamento individual de especialistas não precisa ser informal ou intuitivo. Pesquisadores identificaram significantes diferenças entre "julgamento intuitivo", que tende a ser impreciso [LP92] e "julgamento estruturado", que pode produzir estimativas tão precisas quanto estimativas baseadas em modelos [Jor02].

3.2.2 Wideband Delphi

Wideband Delphi é uma técnica estruturada de estimativa em grupos. A técnica Delphi original foi desenvolvida pela *Rand Corporation* no final da década de quarenta, originalmente como uma maneira de realizar previsões sobre eventos futuros, de onde vem o nome do oráculo grego, localizado no flanco sul do monte Parnassos em Delphi. A técnica Delphi básica consistia em juntar diversos especialistas para criar estimativas independentes e assim, se encontrar pelo tempo necessário para convergir, ou pelo menos concordar, e uma estimativa única. Mais recentemente, a técnica vem sendo utilizada como meio de guiar um grupo de indivíduos informados a um consenso de opinião sobre

um assunto [BAC00]. Participantes são requisitados a fazer uma análise sobre o assunto, individualmente numa rodada inicial, sem consultar outros participantes do exercício. Na primeira rodada, os resultados são colhidos, tabulados e retornados aos participantes para a segunda rodada, durante o qual os participantes são requisitados novamente a fazer uma análise sobre o assunto, mas desta vez com o conhecimento do que foi feito pelos outros participantes na primeira rodada. A segunda rodada normalmente resulta em dados mais alinhados das analises do grupo, apontando para um ponto médio razoável, referente ao assunto em contexto.

Um estudo inicial sobre o uso da técnica Delphi para estimativa de software descobriu que a técnica Delphi básica não era mais precisa do que uma estimativa em grupos menos estruturada. Barry Boehm e seus colegas concluíram que a técnica Delphi genérica era sujeita a muita pressão política e era levada a ser dominada pelos avaliadores mais assertivos do grupo. Conseqüentemente, Boehm e seus colegas estenderam a técnica Delphi básica para o que hoje é chamado de Wideband Delphi [McC06].

Abaixo são descritos os procedimentos básicos:

1. O coordenador Delphi apresenta à cada avaliador a especificação e o formulário de estimativa.
2. Os avaliadores preparam as estimativas iniciais individualmente.
3. O coordenador agenda uma reunião de grupo onde os avaliadores discutem problemas nas estimativas referentes ao projeto.
4. Avaliadores entregam suas estimativas individuais para o coordenador anonimamente.
5. O coordenador prepara um resumo das estimativas em um formulário de iteração e o apresenta para os avaliadores, para que os mesmos possam ver como suas estimativas se comparam com as estimativas dos outros avaliadores.
6. O coordenador agenda uma reunião para discutir as variações nas estimativas.
7. Os avaliadores votam anonimamente se desejam aceitar a estimativa média. Caso algum dos avaliadores discorde, o processo retorna ao terceiro passo.
8. A estimativa final é uma variação criada através da discussão Delphi e a estimativa pontual Delphi é o caso esperado.

Segundo Steve McConnell [McC06], utilizando a técnica Wideband Delphi o erro nas estimativas é reduzido em aproximadamente 40% se comparado a média inicial. McConnell também conclui que a técnnica Wideband Delphi aumenta a precisão na maioria dos casos, e é especialmente útil para evitar erros muito grandes.

Devido a necessidade da técnica Wideband Delphi utilizar reuniões, o tempo consumido por ela é muito grande, sendo assim, uma maneira cara de se estimar um projeto com tarefas muito detalhadas [McC06].

Wideband Delphi é útil para estimar trabalhos em uma nova área de negócios, utilizando uma nova tecnologia ou trabalhando em um novo tipo de software. Ela também é útil caso o projeto dependa muito de especialidades diversas, tais como usabilidade incomum, complexidade algorítmica, performance excepcional, regras de negócios complexas, etc [McC06].

3.3 Decomposição

Decomposição é a pratica de separar uma estimativa em múltiplos pedaços, estimando cada pedaço individualmente, e então, combinar cada estimativa individual em uma estimativa agregada.

Este maneira de se estimar uma tarefa se beneficia de uma propriedade estatística chamada *Lei dos Grandes Números*. O fundamente desta lei está em se você cria uma grande estimativa, a tendência do erro da estimativa é ficar completamente do lado positivo ou do lado negativo, mas se você criar diversas estimativas menores, alguns erros estarão do lado positivo e outros do lado negativo. Sendo assim, os erros nas estimativas tendem a se cancelar a certo nível. Na prática são necessários pelo menos 10 ítens individuais para obter algum benefício da Lei dos Grandes Números, mas mesmo 5 ítens são melhores do que 1 [McC06].

Esta abordagem é comprovada por Lederer e Prassad [LP92], que notaram que a soma das durações das tarefas de um projeto era negativamente correlacionada com excessos de custo e cronograma.

Quão pequenos os pedaços estimados devem ser?

Do ponto de vista de estimativas, desenvolvimento de software é um processo de formar grandes números a partir de pequenas decisões [McC06]. No início de um projeto, uma simples decisão de incluir ou excluir uma funcionalidade pode alterar forma significa-

tiva o esforço e o cronograma de um projeto em uma direção ou outra. Conforme o projeto segue, um maior número de decisões deve ser feito, mas cada uma destas decisões tem um impacto menor sobre o final do projeto no total. A implicação de que o desenvolvimento de software é um processo de refinamento constante é de que quanto mais adiante dentro de um projeto você está, mais decompostas as estimativas podem ser [McC06]. Os limites no número de ítens a se estimar são mais práticos do que teóricos, variando muito entre projetos, e devem se adaptar a sua realidade.

3.3.1 Decomposição por Work Breakdown Structure

Um padrão de longa data no desenvolvimento de hardware e software, WBS é a maneira de organizar elementos do projeto em uma hierarquia que simplifica as tarefas de estimar custo e controle. Ela ajuda a determinar exatamente quais custos estão sendo estimados [BAC00]. Basicamente, se probabilidades são determinadas a custos associados a cada elemento da hierarquia, o valor total esperado pode ser determinado analisando os custos do projeto, de baixo para cima. Experiência é associada a este método na especificação dos elementos mais úteis dentro da estrutura e quais as probabilidades associadas a cada elemento.

Uma WBS de software consiste atualmente de duas hierarquias, uma representando o produto de software e a outra representando as atividades necessárias para construção do produto. A hierarquia do produto descreve a estrutura fundamental do software, demonstrando como diversos elementos do software se encaixam no sistema por inteiro.

A hierarquia de atividade indica as atividades que devem ser associadas com cada elemento de software.

Algumas vezes, trabalho despercebido se esconde um funções esquecidas. Algumas vezes ele se esconde em forma de tarefas esquecidas [McC06]. Decompondo um projeto utilizando uma WBS baseada em atividades ajuda a evitar tarefas esquecidas. Ele também ajuda a melhorar o entendimento sobre o tamanho do projeto a ser estimado, baseado em projetos passados. A comparação de uma WBS de um projeto novo com uma de um projeto antigo pode ajudar a aguçar sua avaliação sobre quais partes são menores ou quais partes são maiores [McC06].

Além de ajudar com a estimativa, outra grande utilidade da WBS é a contabilidade de custo. Cada elemento da WBS pode ser associado ao seu próprio custo, permitindo assim, a analise de tempo gasto em cada tarefa ou elemento do projeto, informação que

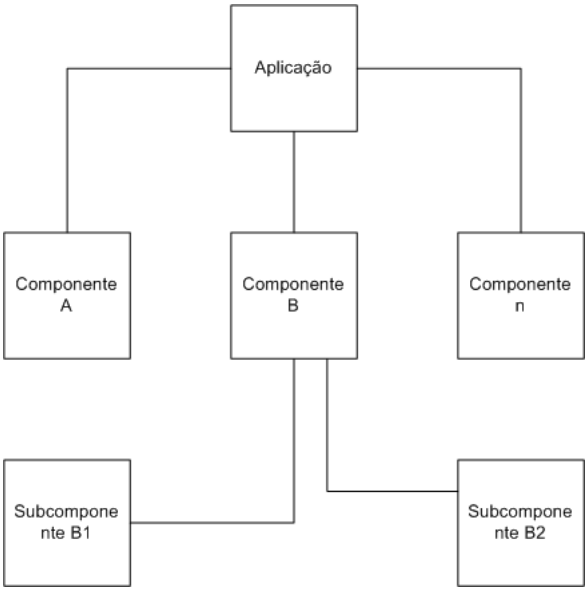


Figura 3.1: Hierarquia do produto

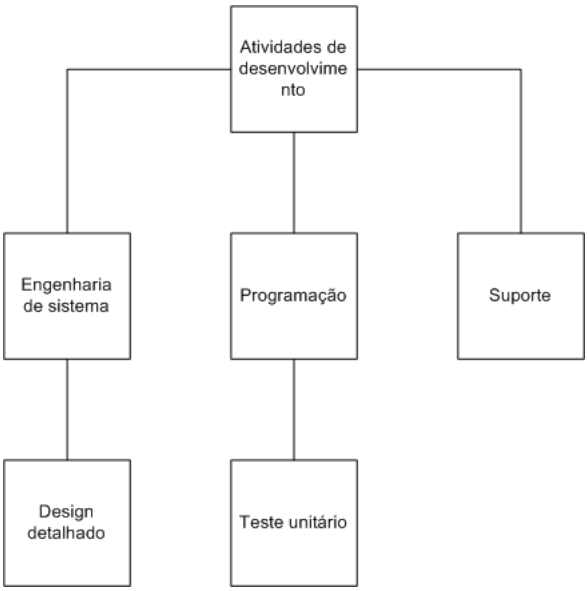


Figura 3.2: Hierarquia de atividades

pode ser sumarizada para controle da verba do projeto.

Finalmente, se uma organização utiliza de uma forma consistente uma WBS padrão para todos os projetos, com o tempo uma base de dados valiosa refletindo custos de projeto será criada. Estes dados poderão ser utilizados para desenvolver um modelo de estimativa de custo ajustado para a experiência e praticas da organização.

3.4 Modelos de custo de software

A pesquisa no campo de modelagem de custo de software começou com o extensivo estudo de E. Nelson, da *Systems Development Corporation* (SDC) em 1965 sobre 104 atributos implementados em 169 projetos de software, e isto evoluiu posteriormente, no final da década de sessenta e inicio da década de setenta, para alguns modelos parciais que serão citados no decorrer do texto.

No final da década de setenta, modelos mais robustos foram produzidos, como o modelo SLiM, de L. Putnam e W. Myers [PM92], o modelo Checkpoint, de C. Jones [Jon91], o modelo PRICE-S, de R. Park [Par88], SEER-SEM, de R. Jensen [Jen83] e o modelo COCOMO, de B. Boehm [Boe81], entre outros.

A maioria destes pesquisadores iniciou seus trabalhos na área de desenvolvimento de modelos de custo de software aproximadamente ao mesmo tempo, e todos eles encontraram o mesmo problema: conforme um software crescia em tamanho e importância, ele também crescia em complexidade, fazendo com que a tarefa de prever o custo de desenvolvimento de um software se tornasse cada vez mais difícil.

Assim como qualquer outro campo, o campo de modelagem de custos de software possui suas próprias armadilhas. Assim como a engenharia de software, a modelagem de custo de software está em constante mudança, o que tornou difícil o desenvolvimento de modelos paramétricos que obtivessem alta precisão para o desenvolvimento de software em todos os domínios.

Sendo assim, os custos de desenvolvimento de software continuam a crescer e profissionais continuamente expressam suas preocupações sobre a inabilidade de prever com precisão os custos envolvidos. Um dos objetivos mais importantes da comunidade de engenharia de software foi o desenvolvimento de modelos úteis que construtivamente expliquem o ciclo de vida de desenvolvimento de softwares e prevejam com precisão o custo de um software.

Na categoria de técnicas baseadas em modelos se enquadram diversas técnicas derivadas inteiramente ou parcialmente de um modelo que descreve diversos dos aspectos do projeto em questão.

O modelo deve utilizar como entrada diversos aspectos conhecidos do projeto, tais como tamanho, recursos, complexidade, ambiente e limitações. Baseado em tais entradas, o modelo deve gerar previsões sobre o projeto, dentre elas o esforço necessário, organização

de recursos, estimativas de custo e defeitos, entre outros, dependendo do modelo utilizado.

Muitos dos modelos presentes nesta categoria são modelos proprietários, logo, não podem ser comparados em termos de estrutura. Teoria ou experimentação determina a forma funcional destes modelos.

Diversas técnicas de estimativa foram analisadas neste contexto, tais como SLiM, Checkpoint, PRICE-S, ESTIMACS, SEER-SEM, SELECT Estimator e COCOMO2, mas não serão abordadas diretamente neste trabalho.

4 Conceitos de estatística

4.1 Regressão

Análise de regressão refere-se à descrição e a quantificação da relação entre uma dada variável, em geral chamada de dependente, e uma ou mais variáveis, em geral chamadas independentes [Mad01].

O termo regressão foi concebido pelo inglês Sir Francis Galton, que estudava a correlação entre a altura das crianças e a altura dos pais. Ele observou que, embora pais altos tivessem filhos altos e pais baixos tivessem filhos baixos, havia uma tendência para que a altura das crianças convergisse à media. Há, portanto, "uma regressão da altura das crianças em direção à média" [Mad01].

Quando possuímos apenas uma variável independente, como no caso de Galton, onde apenas a altura dos pais era correlacionada com a altura dos filhos, a regressão é denominada simples. Em casos onde diversas variáveis independentes são correlacionadas com uma variável dependente, temos o que é chamado de regressão múltipla, assunto que será discutido mais adiante.

4.1.1 Regressão simples

Como vimos anteriormente, é denominada regressão simples a análise da correlação entre uma variável dependente e uma variável independente. Há vários objetivos em se estudar estas relações e dentre elas, duas serão analisadas neste trabalho:

- Prever o valor de uma variável dependente para um conjunto de variáveis independentes.
- Examinar a significância de variáveis independentes sobre a variável dependente.

A seguir, iremos analisar os métodos utilizados na regressão simples. Nos exemplos a seguir, denotaremos a variável dependente por y e a variável independente por x.

A relação entre y e x é denotada por

$$y = f(x)$$

onde *f(x)* é uma função de x.

Precisamos notar a distinção entre uma relação desterminística ou matemática e uma relação estatística, que não fornece um valor exato de y dado um valor de x [Mad01].

A análise de regressão trata de relações estatísticas, onde os valores de y para diferentes valores de x não podem ser determinados com exatidão, mas podem ser descritos probabilisticamente.

Exemplo Ilustrativo

Considere o seguinte exemplo, citado por G.S. Maddala [Mad01]. Suponha a relação entre vendas y e gastos com publicidade x seja

$$y = 2500 + 100x - x^2$$

Esta é uma relação determinística, onde as vendas podem ser determinadas com exatidão baseadas nos gastos com publicidade. Esta relação é dada da seguinte forma:

x	y
0	2500
20	4100
50	5000
100	2500

Tabela 4.1: Relação determinística entre vendas e gastos com propaganda [Mad01]

Suponha, porém, que a relação entre vendas y e gastos com propagandas x seja

$$y = 2500 + 100x - x^2 + u$$

onde $u = +500$ com probabilidade $\frac{1}{2}$ e -500 com probabilidade $\frac{1}{2}$.

Desta forma, os valores de y para diferentes valores de x não podem ser determinados com exatidão, mas podem ser descritos probabilisticamente. Se, por exemplo, os gastos com propaganda forem 50, as vendas serão 5500 com probabilidade $\frac{1}{2}$ e 4500 com probabilidade $\frac{1}{2}$. Agora os valores de y para diferentes valores de x são os seguintes:

x	y
0	2000 ou 3000
20	3600 ou 4600
50	4500 ou 5500
100	2000 ou 3000

Tabela 4.2: Relação estatística entre vendas e gastos com propaganda [Mad01]

Se o termo de erro u tem uma distribuição contínua, digamos uma distribuição normal com média 0 e variância 1, então, para cada valor de x, teremos uma distribuição normal para y e o valor observado de y poderá ser qualquer observação desta distribuição.

Podemos demonstrar esta distribuição utilizando a relação

$$y = 2 + x + u$$

que pode ser vista na Figura 4.1

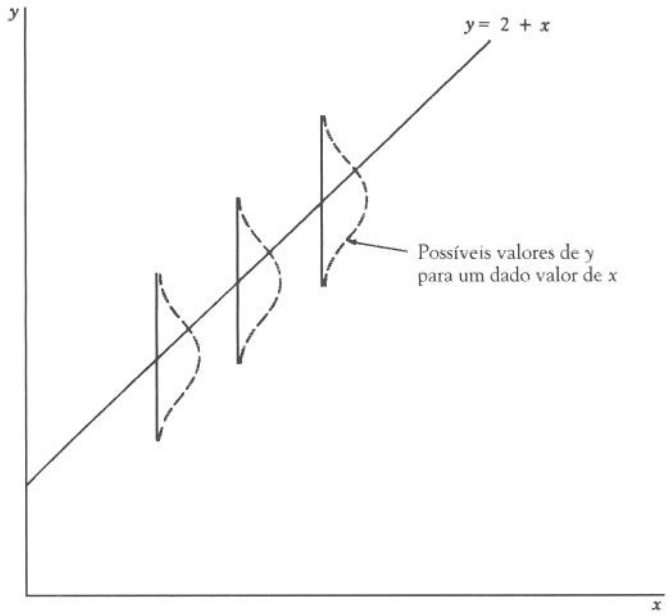


Figura 4.1: Uma correlação estatística [Mad01]

Levando em consideração o exemplo acima, temos uma função linear, que pode ser definida como

$$f(x) = \alpha + \beta x$$

e assumiremos que esta relação é estatística

$$f(x) = \alpha + \beta x + u$$

onde u, chamado de erro ou distúrbio, tem uma distribuição de probabilidade conhecida. Este termo possui diversas fontes, dentre elas, podemos destacar:

1. Elementos não-previsíveis de aleatoriedade nas respostas humanas.
2. Efeito de um grande número de variáveis omitidas.
3. Erro de mensuração em y.

O objetivo de uma análise de regressão é calcular estimativas dos parâmetros desconhecidos α e β, dadas n observações de x e y.

Inicialmente devemos trabalhar com alguns pressupostos sobre os termos de erro u:

1. *Média zero.* $E(u_i) = 0$ para todo i.
2. *Variância comum.* $var(u_i) = \sigma^2$ para todo i.
3. *Independência.* u_i e u_j são independentes para todo $i \neq j$.
4. *Independência de* x_j. u_i e x_j são independentes para todo i e j.
5. *Normalidade.* u_i é normalmente distribuído para todo i.

Estas são hipóteses iniciais. Algumas serão relaxadas posteriormente.

Discutiremos a seguir, alguns dos métodos utilizados na análise de regressão.

4.1.1.1 Método dos momentos

Levando em consideração os pressupostos 1 e 2 citados anteriormente, no método dos momentos, substituímos estas condições por suas contrapartes amostrais [Mad01].

Sejam $\hat{\alpha}$ e $\hat{\beta}$ os estimadores de α e β, respectivamente. A contraparte amostral de u_i é o erro estimado $\hat{u}_i$, definido por

$$\hat{u}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$$

As duas equações para se determinar $\hat{\alpha}$ e $\hat{\beta}$ são obtidas pela substituição dos pressupostos populacionais por suas contrapartes amostrais [Mad01]:

Hipótese sobre a População	Contraparte Amostral
$E(u) = 0$	$\frac{1}{n}\sum \hat{u}_i = 0$ ou $\sum \hat{u}_i = 0$
$cov(x, u) = 0$	$\frac{1}{n}\sum x_i\hat{u}_i = 0$ ou $\sum x_i\hat{u}_i = 0$

Substituindo $\hat{u}_i$ temos as seguintes equações

$$\sum \hat{u}_i = 0 \text{ ou } \sum (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0$$
$$\sum x_i\hat{u}_i = 0 \text{ ou } \sum x_i(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0$$

que podem ser escritas como

$$\sum y_i = n\hat{\alpha} + \hat{\beta}\sum x_i$$
$$\sum x_i y_i = \hat{\alpha}\sum x_i + \hat{\beta}\sum x_i^2$$

Resolvendo estas equações, encontramos os valores de $\hat{\alpha}$ e $\hat{\beta}$. Estas equações também são chamadas de "equações normais".

O método dos mínimos quadrados, descrito a seguir, se baseia no princípio da escolha de $\hat{\alpha}$ e $\hat{\beta}$ de forma que $\sum \hat{u}_i^2$ seja mínimo. Isto é, a soma dos quadrados dos erros previstos é mínima. Com este método achamos as mesmas estimativas de α e β, uma vez que obtemos as mesmas equações normais.

4.1.1.2 Método dos mínimos quadrados

O método dos mínimos quadrados é o motor da análise estatística moderna; apesar de suas limitações, acidentes ocasionais e poluição incidental, ele e suas numerosas variações, extensões e suas conveniências correlatas são o cerne da análise estatística, e são valorizados por isto [Mad01].

O método dos mínimos quadrados requer que escolhamos $\hat{\alpha}$ e $\hat{\beta}$ como estimadores

de α e β, respectivamente, de forma que

$$Q = \sum (Y - i - \hat{\alpha} - \hat{\beta} x_i)^2$$

seja mínimo. Q também é a soma dos quadrados dos erros previstos quando estimamos y_i dado x_i e a equação de regressão estimada.

O procedimento de minimizar o fator Q na equação em relação a $\hat{\alpha}$ e $\hat{\beta}$ necessita que calculemos as primeiras derivadas em relação a $\hat{\alpha}$ e $\hat{\beta}$ e as igualemos a zero. Este procedimento nos dá

$$\frac{\partial Q}{\partial \hat{\alpha}} = 0 \Rightarrow \sum 2(y_i - \hat{\alpha} - \hat{\beta} x_i)(-1) = 0$$

ou

$$\sum y_i = n\hat{\alpha} + \hat{\beta} \sum x_i$$

ou

$$\overline{y} = \hat{\alpha} + \hat{\beta} \overline{x} \tag{4.1}$$

e

$$\frac{\partial Q}{\partial \hat{\alpha}} = 0 \Rightarrow \sum 2(y_i - \hat{\alpha} - \hat{\beta} x_i)(-x_i) = 0$$

ou

$$\sum y_i x_i = \hat{\alpha} \sum x_i + \hat{\beta} \sum x_i^2 \tag{4.2}$$

As Equações 4.1 e 4.2 são chamadas equações normais [Mad01]. Substituindo o valor de $\hat{\alpha}$, temos

$$\begin{aligned} \sum y_i x_i &= \sum x_i(\overline{y} - \hat{\beta}\overline{x}) + \hat{\beta} \sum x_i^2 \\ &= n\overline{x}(\overline{y} - \hat{\beta}\overline{x}) + \hat{\beta} \sum x_i^2 \end{aligned}$$

Definindo

$$S_{yy} = \sum (y_i - \overline{y})^2 = \sum y_i^2 - n\overline{y}^2$$
$$S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - n\overline{xy}$$
$$S_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2$$

Então a equação pode ser escrita da seguinte maneira

$$\hat{\beta} S_{xx} = S_{xy} \text{ ou } \hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

Portanto os estimadores de mínimos quadrados para α e β são [Mad01]

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \text{ e } \hat{\alpha} = \overline{y} - \hat{\beta} x_i$$

Os resíduos estimados são [Mad01]

$$\hat{u}_i = y_i - \hat{\alpha} - \hat{\beta}\overline{x}$$

As duas equações normais mostram que estes resíduos satisfazem as equações [Mad01]

$$\sum \hat{u}_i = 0$$

e

$$\sum x_i \hat{u}_i = 0$$

A soma dos quadrados dos resíduos (denotada por SQR) é dada por [Mad01]

$$
\begin{aligned}
SQR &= \sum (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \\
&= \sum [y_i - \overline{y} - \hat{\beta}(x_i - \overline{x})]^2 \\
&= \sum (y_i - \overline{y})^2 + \hat{\beta}^2 \sum (x_i - \overline{x})^2 - 2\hat{\beta} \sum (y_i - \overline{y})(x_i - \overline{x}) \\
&= S_{yy} + \hat{\beta}^2 S_{xx} - 2\hat{\beta} S_{xy}
\end{aligned}
$$

Mas $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$. Logo, temos

$$SQR = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = S_{yy} - \hat{\beta} S_{xy}$$

S_{xy} é geralmente denotado por SQT e $\hat{\beta} S_{xy}$ como SQE [Mad01]. Assim,

$$\text{SQT (total)} = \text{SQE} + \text{SQR}$$

A proporção da soma total dos quadrados explicados é denotada por r_{xy}^2, onde r_{xy} é chamado de coeficiente de correlação. Logo $r_{xy}^2 = SQE/SQT$ e $1 - r_{xy}^2 = SQR/SQT$.

O termo r_{xy}^2 é chamado de coeficiente de determinação, e é distribuído entre 0 e 1 para qualquer regressão. Se r_{xy}^2 está próximo de 0, a variável x explica muito pouco das variações de y. Se r_{xy}^2 está próximo de 1, a variável x explica a maior parte das variações de y.

O coeficiente de determinação r_{xy}^2 é dado por

$$r_{xy}^2 = \frac{SQE}{SQT} = \frac{SQT - SQR}{SQT} = \frac{\hat{\beta} S_{xy}}{S_{yy}}$$

Exemplo Ilustrativo

Considere o seguinte exemplo, citado por G. S. Maddala [Mad01]. Suponha os seguintes dados para 10 trabalhadores expressos na Tabela 4.3, onde

$$x = \text{horas de produção}$$
$$y = \text{produção}$$

Tendo em vista tais dados, queremos determinar a relação entre produção e horas

Observação	x	y	x^2	y^2	xy
1	10	11	100	121	110
2	7	10	49	100	70
3	10	12	100	144	120
4	5	6	25	36	30
5	8	10	64	100	80
6	8	7	64	49	56
7	6	9	36	81	54
8	7	10	49	100	70
9	9	11	81	121	99
10	10	10	100	100	100
Total	80	96	668	952	789

Tabela 4.3: Horas de trabalho e produção [Mad01].

de trabalho. Assim temos

$$\begin{aligned}
\overline{x} &= \frac{80}{10} = 8 \\
\overline{y} &= \frac{96}{10} = 9,6 \\
S_{xx} &= 668 - 10(8)^2 = 668 - 640 = 28 \\
S_{xy} &= 789 - 10(8)(9,6) = 789 - 768 = 21 \\
S_{yy} &= 952 - 10(9,6)^2 = 952 - 921,6 = 30,4 \\
r_{xy} &= \frac{21}{\sqrt{28(30,4)}} \approx \frac{21}{29} = 0,724 \text{ ou } r_{xy}^2 = 0,52 \\
\hat{\beta} &= \frac{S_{xy}}{S_{xx}} = \frac{21}{28} = 0,75 \\
\hat{\alpha} &= \overline{y} - \hat{\beta}\overline{x} = 9,6 - 0,75(0,8) = 3,6
\end{aligned}$$

Logo, a regressão de y em x é

$$y = 3,6 + 0,75x$$

4.1.1.3 Inferência estatística no modelo de regressão linear

Para encontrar os estimados mínimos quadrados de α e β não é necessário assumir qualquer distribuição de probabilidade particular para os erros u_i, mas para calcular estimadores intervalares dos parâmetros precisamos assumir que os erros de u_i têm uma

distribuição normal [Mad01]. Os estimadores mínimos quadrados:

1. São não-tendenciosos.
2. Têm a menor variância dentre a classe de estimadores lineares não-tendenciosos.

Estas propriedades são válidas mesmo quando os erros u_i não possuírem uma distribuição normal, desde que os seguintes pressupostos estabelecidos anteriormente sejam satisfeitos [Mad01]:

1. *Média zero.* $E(u_i) = 0$ para todo i.
2. *Variância comum.* $var(u_i) = \sigma^2$ para todo i.
3. *Independência.* u_i e u_j são independentes para todo $i \neq j$.
4. *Independência de* x_j. u_i e x_j são independentes para todo i e j.

Estabelecendo o pressuposto adicional de que os erros de u_i são normalmente distribuídos, podemos calcular intervalos de confiança para α e β. Tendo em vista que a variância do erro σ^2 não é conhecida, temos que estimá-la [Mad01]. Se SQR é a soma dos quadrados residuais, então

$$\hat{\sigma}^2 = \frac{SQR}{n-2} \text{ é um estimador não-viesado (neutro) de } \sigma^2$$

Também

$\frac{SQR}{\sigma^2}$ tem uma distribuição χ^2 ou "chi-quadrado" com (n - 2) graus de liberdade.

Neste ponto é necessário fazer inferências sobre α e β. Para este propósito, utilizamos a distribuição t de student. Se temos duas variáveis, x_1 com distribuição normal, média 0 e variância 1, e x_2 com distribuição "chi-quadrado" com k graus de liberdade, e x_1 e x_2 são independentes, então

$$x = \frac{x_1}{\sqrt{x_2/k}} = \frac{\text{normal padrão}}{\sqrt{\chi^2 \text{independente médio}}}$$

tem uma distribuição t com k graus de liberdade [Mad01].

Neste caso $(\hat{\beta}-\beta)/\sqrt{\sigma^2/S_{xx}}$ possui uma distribuição normal, média 0 e variância 1 e SQR/σ^2 possui uma distribuição "chi-quadrado" com $n-2$ graus de liberdade e as duas distribuições são independentes. Definimos a razão

$$\frac{\hat{\beta}-\beta}{\sqrt{\sigma^2/S_{xx}}}\Big/\sqrt{\frac{SQR}{(n-2)\sigma^2}}$$

que tem uma distribuição t com $(n-2)$ graus de liberdade. σ^2 é cancelado e, escrevendo $SQR/(n-2)$ como σ^2, temos como resultado que, $(\hat{\beta}-\beta)/\sqrt{\sigma^2/S_{xx}}$ tem uma distribuição t com $(n-2)$ graus de liberdade [Mad01]. Como σ^2 não é conhecido, utilizamos um estimador não-tendencioso $SQR/(n-2)$. Desta forma, $\hat{\sigma}^2/S_{xx}$ é a variância estimada de $\hat{\beta}$ e sua raiz quadrada é chamada de *erro padrão*, denotado por $EP(\hat{\beta})$ [Mad01]. Podemos seguir um procedimento similar para $\hat{\alpha}$, substituindo $\hat{\sigma}^2$ por σ^2 na variância de $\hat{\alpha}$ e tiramos sua raiz quadrada para calcular o erro padrão $EP(\hat{\alpha})$ [Mad01].

Exemplo Ilustrativo

Continuando o exemplo citado por G. S. Maddala [Mad01] anteriormente, obtivemos a equação de regressão de y em x como

$$\hat{y} = 3,6 + 0,75x$$

Podemos calcular agora os erros padrões de $\hat{\alpha}$ e $\hat{\beta}$ da seguinte maneira:

1. Calculando as variâncias de $\hat{\alpha}$ e $\hat{\beta}$ em termos de $\hat{\sigma}^2$.
2. Substituindo σ^2 por $\hat{\sigma}^2 = SQR/(n-2)$.
3. Extraindo a raiz quadrada das expressões resultantes.

Temos assim

$$V(\hat{\alpha}) = \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}) = \sigma^2(\frac{1}{10} + \frac{64}{28}) = 2,39\sigma^2$$
$$V(\hat{\beta}) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{28} = 0,036\sigma^2$$
$$\sigma^2 = \frac{1}{n-2}(S_{yy} - \frac{S_{xy}^2}{S_{xx}}) = \frac{1}{8}[30,4 - \frac{(21)^2}{28}] = 1,83$$
$$EP(\hat{\alpha}) = \sqrt{(2,39)(1,83)} = 2,09$$
$$EP(\hat{\beta}) = \sqrt{(0,036)(1,83)} = 0,256$$

Em posse dos erros padrões, podemos estimar os intervalos de confiança para α, β e σ^2 [Mad01]. Como $(\hat{\alpha} - \alpha)/EP(\hat{\alpha})$ e $(\hat{\beta} - \beta)/EP(\hat{\beta})$ têm distribuições t com $(n-2)$ graus de liberdade, utilizando a tabela da distribuição t (Apêndice A) com $n - 2 = 8$ graus de liberdade temos

$$P[-2,306 < \frac{\hat{\alpha} - \alpha}{EP(\hat{\alpha})} < 2,306] = 0,95$$

e

$$P[-2,306 < \frac{\hat{\beta} - \beta}{EP(\hat{\beta})} < 2,306] = 0,95$$

Isto nos dá os intervalos de confiança para α e β [Mad01]. Substituindo os valores de $\hat{\alpha}$, $\hat{\beta}$, $EP(\hat{\alpha})$ e $EP(\hat{\beta})$ e simplificando, temos o limite do intervalo de confiança a 95% para α e β como (-1,22;8,42) para α e (0,16;1,34) para β [Mad01]. Note que os limites do intervalo de confiança para α a 95% são $\hat{\alpha} \pm 2,306EP(\hat{\alpha})$ e $\hat{\beta} \pm 2,306EP(\hat{\beta})$ para β.

podemos produzir intervalos menores reduzindo o coeficiente de confiança. Por exemplo, os limites de confiança a 80% para β são

$$\hat{\beta} \pm 1,397EP(\hat{\beta}) \text{ pois } P(-1,397 < t < 1,397) = 0,80$$

da tabela t com 8 graus de liberdade. Desta forma encontramos os limites de confiança para β como

$$0,75 \pm 1,397(0,256) = (0,39; 1,11)$$

4.1.1.4 Previsão com o modelo de regressão simples

A equação de regressão estimada $\hat{y} = \hat{\alpha} + \hat{\beta}$ é utilizada para prever o valor de y dados os valores de x.

Seja x_0 um dado valor de x, podemos prever o correspondente valor de y, y_0, através de

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0$$

O verdadeiro valor de y_0 é dado por

$$y_0 = \alpha + \beta x_0 + u_0$$

onde u_0 é o termo de erro. Logo o erro previsto é

$$\hat{y}_0 - y_0 = (\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)x_0 - u_0$$

Como $E(\hat{\alpha} - \alpha) = 0$, $E(\hat{\beta} - \beta) = 0$ e $E(u_0) = 0$, temos $E(\hat{y}_0 - y_0) = 0$.

A variância do erro previsto é

$$\begin{aligned} V(\hat{y}_0 - Y_0) &= V(\hat{\alpha} - \alpha) + x_0^2 V(\hat{\beta} - \beta) + 2x_0 cov(\hat{\alpha} - \alpha, \hat{\beta} - \beta) + V(u_0) \\ &= \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}) + \sigma^2 \frac{x_0^2}{S_{xx}} - 2_{x_0}\sigma^2 \frac{\overline{x}}{S_{xx}} + \sigma^2 \\ &= \sigma^2[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}] \end{aligned}$$

Portanto, a variância aumenta quanto mais distante o valor de x_0 está de $\overline{x}$, que é a média das observações, valor utilizado para o cálculo de α e β [Mad01].

Se x_0 estiver dentro do intervalo das observações amostrais em x, podemos chamá-lo de previsão intra-amostral, e se x_0 estiver fora do intervalo das observações amostrais, podemos chamá-lo de previsão extra-amostral [Mad01].

Um fator a mais a se notar é que não é apropriado obter previsões muito além do intervalo de observações, já que as equações de regressão foram estimadas com base em um conjunto de valores x, que pode variar de uma maneira diferente fora o intervalo

amostral.

Exemplo Ilustrativo

Como ilustração, considere o 2 citado por G.S. Maddala [Mad01], onde é dada a seguinte função de consumo estimada a partir de 12 observações.

$$y = 10 + 0,9x$$

onde y denota gastos com consume e x denota a renda disponível.

Temos $\sigma^2 = 0,01$, $\overline{x} = 200$ e $S_{xx} = 4000$. Dado $x_0 = 250$, nossa previsão de y_0 é

$$\hat{y}_0 = 10 + 0,9(250) = 235$$
$$EP(\hat{y}_0 = \sqrt{0,01(1 + \frac{1}{12} + \frac{2500}{4000})} = 0,131)$$

Como $t = 2,228$ das tabelas t com 10 graus de liberdade, o intervalo de confiança a 95% para y_0 é $235 \pm 2,228(0,131) = 235 \pm 0,29$, isto é $(234,71; 235,29)$ [Mad01].

4.1.1.5 Pontos discrepantes

É muito comum que as estimativas dos parâmetros de regressão sejam influenciadas por algumas poucas observações extremas ou discrepantes. Este problema pode ser detectado se analisarmos os resíduos da equação de regressão estimada, e deve sempre acompanhar toda equação estimada, pois nos permitirá justificar o estabelecimento dos seguintes pressupostos [Mad01]:

1. A variância do erro $V(u_i) = \sigma^2$ para todo i.
2. Os termos de erro são serialmente independentes.
3. A forma funcional da regressão é linear ou não.

Estamos interessados em detectar algumas observações discrepantes utilizando a análise dos resíduos. Uma observação discrepante, ou ponto discrepante é uma observação que está muito distante do restante das observações. Ela costuma ser gerada por algum fator incomum. Entretanto, quando utilizamos o método dos mínimos quadrados, esta única observação pode produzir mudanças substanciais na equação de regressão estimada.

No caso de uma regressão simples, podemos detectar estes pontos simplesmente pela tabulação dos dados. No caso da regressão múltipla, tal tabulção não é possível, de forma que temos de analisar os resíduos $\hat{u}_i$ [Mad01]. Um bom exemplo de que uma representação simples da equação de regressão com os erros padrões associados não nos fornece todas as informações é demonstrado em [Ans73]. Quatro conjuntos de dados fornecem a mesma equação de regressão.

Conjunto de Dados		1-3	1	2	3	4	4
Variável		x	y	y	y	x	y
Observação	1	10	8,04	9,14	7,46	8	6,58
	2	8	6,95	8,14	6,77	8	5,76
	3	13	7,58	8,74	12,74	8	7,71
	4	9	8,81	8,77	7,11	8	8
	5	11	8,33	9,26	7,81	8	8,47
	6	14	9,96	8,1	8,84	8	7,04
	7	6	7,24	6,13	6,08	8	5,25
	8	4	4,26	13,1	5,39	19	12,5
	9	12	10,84	9,13	8,15	8	5,56
	10	7	4,82	7,26	6,42	8	7,91
	11	5	5,68	4,74	5,73	8	6,89

Tabela 4.4: Quatro conjuntos de dados [Ans73]

Para os quatro conjuntos de dados, temos as seguintes estatísticas:

$$\begin{array}{ccc} n = 11 & \overline{x} = 9,0 & \overline{y} = 7,5 \\ S_{xx} = 110,0 & S_{yy} = 41,25 & S_{xy} = 55,0 \end{array}$$

A equação de regressão é

$$\begin{aligned} \hat{y} &= 3,0 + 0,5x \\ r^2 &= 0,667 \\ SQE &= 27,50(1g.l.) \\ SQR &= 13,75(9g.l.) \end{aligned}$$

Embora as equações de regressão sejam idênticas, os quatro conjuntos de dados exibem características muito diferentes, como podemos ver na Figura 4.2.

O conjunto de dados 1 na Figura 4.2 (i) não mostra nenhum problema especial. A Figura 4.2 (ii) mostra que a reta de regressão não deve ser linear. A Figura 4.2 (iii)

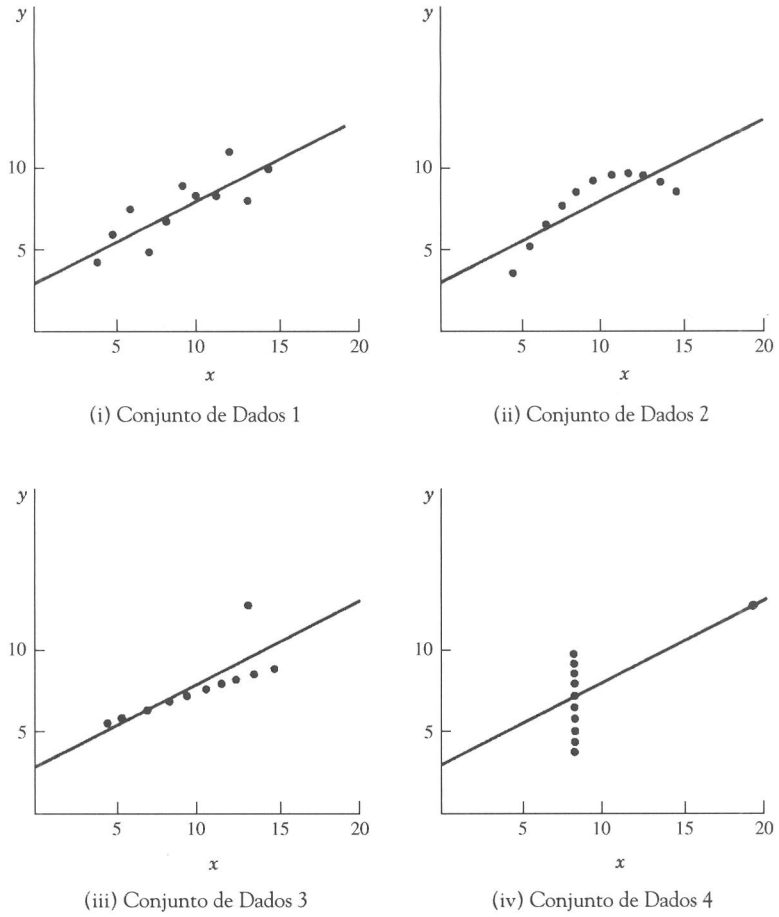


Figura 4.2: Linhas de regressão para quatro conjuntos de dados [Ans73]

mostra como um único ponto deturpou levemente a reta de regressão. Se omitíssemos esta única observação, teríamos uma reta de regressão ligeiramente diferente. A Figura 4.2 (iv) mostra como um ponto pode produzir uma situação inteiramente diferente. Se esta observação fosse omitida, teríamos uma reta de regressão vertical.

Exemplo Ilustrativo

Para demonstrar como podemos utilizar a análise do resíduo da regressão para encontrar pontos discrepantes, utilizaremos um exemplo citado por G. S. Maddala [Mad01]. Este exemplo consiste na estimativa da função de consumo dos Estados Unidos para o período de 1929 a 1984. A Tabela 4.5 nos dá os dados da renda disponível *Per Capita* e os gastos *Per Capita* com consumo. Os resultados da regressão são os seguintes:

Ano	C	Y	Ano	C	Y	Ano	C	Y
1929	1765	1883	1957	2416	2660	1977	3924	4280
1933	1356	1349	1958	2400	2645	1978	4057	4441
1939	1678	1754	1959	2487	2709	1979	4121	4512
1940	1740	1847	1960	2501	2709	1980	4093	4487
1941	1826	2083	1961	2511	2742	1981	4131	4561
1942	1788	2354	1962	2583	2813	1982	4146	4555
1943	1815	2429	1963	2644	2865	1983	4303	4670
1944	1844	2483	1964	2751	3026	1984	4490	4941
1945	1936	2416	1965	2868	3171			
1946	2129	2353	1966	2979	3290			
1947	2122	2212	1967	3032	3389			
1948	2129	2290	1968	3160	3493			
1949	2140	2257	1969	3245	3564			
1950	2224	2392	1970	3277	3665			
1951	2214	2415	1971	3355	3752			
1952	2230	2441	1972	3511	3860			
1953	2277	2501	1973	3623	4080			
1954	2278	2483	1974	3566	4009			
1955	2384	2582	1975	3609	4051			
1956	2410	2453	1976	3774	4158			

Tabela 4.5: Gastos com consumo pessoal e renda pessoal disponível [Mad01]

$$\hat{C} = -24,944 + 0,911Y$$
$$r^2 = 0,9823$$

O próximo passo é examinar os resíduos, os quais são apresentados na Tabela 4.6 [Mad01]. Pode-se facilmente notar os grandes resíduos negativos para as observações 6,7,8 e 9. Estas observações são pontos discrepantes, pois correspondem aos anos de guerra entre 1942 e 1945, durante os quais, controles severos foram impostos nos gastos com consumo. Podemos desta forma descartar estar observações e reestimar a equação [Mad01].

$$\hat{C} = 85,725 + 0,885Y$$
$$r^2 = 0,9975$$

Observação	Resíduo	Observação	Resíduo	Observação	Resíduo
1	75,0	17	24,2	33	24,1
2	152,4	18	41,6	34	-35,9
3	105,5	19	57,4	35	-37,1
4	82,8	20	18,8	36	20,5
5	-46,1	21	18,4	37	-67,9
6	-330,9	22	16,0	38	-60,2
7	-372,2	23	44,8	39	-55,4
8	-392,4	24	58,8	40	12,1
9	-239,4	25	38,7	41	51,0
10	11,0	26	46,0	42	37,4
11	132,4	27	59,7	43	36,7
12	68,4	28	20,1	44	31,5
13	109,4	29	5,0	45	2,1
14	70,5	30	7,6	46	22,5
15	39,5	31	-29,5	47	74,8
16	31,8	32	3,7	48	15,0

Tabela 4.6: Resíduos para função de consumo estimada [Mad01]

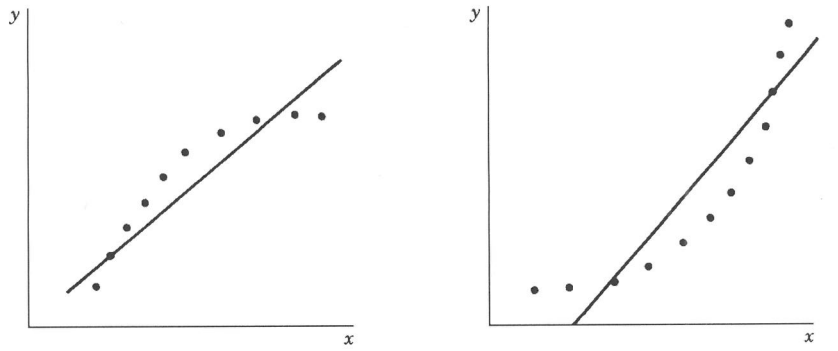
Observação	Resíduo	Observação	Resíduo	Observação	Resíduo
1	12,4	20	-24,2	35	-52,1
2	76,1	21	-24,4	36	8,3
3	39,6	22	-27,2	37	-74,5
4	19,2	23	3,2	38	-68,6
5	-103,7	24	17,2	39	-62,8
10	-39,7	25	-2,0	40	7,5
11	78,1	26	7,1	41	49,5
12	16,1	27	22,1	42	40,0
13	56,3	28	-13,4	43	41,1
14	20,8	29	-24,8	44	35,2
15	-9,6	30	-19,1	45	7,7
16	-16,6	31	-53,8	46	28,0
17	-22,7	32	-17,8	47	83,2
18	-5,8	33	4,3	48	30,3
19	12,6	34	-53,1		

Tabela 4.7: Resíduos para função de consumo estimada omitindo os anos de guerra [Mad01]

Anteriormente o intercepto não era significativamente diferente de zero. Agora ele é significativamente positivo. Além disto, a estimativa da propensão marginal ao consume é significativamente menor. Os resíduos estimados desta equação são apresentados na Tabela 4.7, onde não podemos observar nenhum resíduo excepcionalmente grande, nem longas seqüencias de resíduos positivos ou negativos, como na Tabela 4.6 [Mad01].

4.1.1.6 Formas funcionais alternativas para equações de regressão

Vimos anteriormente, com referência na Figura 4.2 (ii), que algumas vezes a correlação entre y e x pode não ser linear. Neste caso, devemos assumir uma forma funcional apropriada para esta relação [Mad01]. Há várias formas funcionais passíveis de serem utilizadas e, após algumas transformações das variáveis, podem ser introduzidas à estrutura usual de regressão linear.



(a) y aumenta mais lentamente do que x (a) y aumenta mais rapidamente do que x

Figura 4.3: Regressão linear não apropriada ao conjunto de dados [Mad01]

Como exemplo, podemos observar os dados representados pelos pontos expressos na Figura 4.3 (a), onde y aumenta mais lentamente do que x. Uma forma funcional possível é $y = \alpha + \beta \log x$. Isto é chamado de forma *semilog*, já que ela envolve o logaritmo para apenas uma das duas variáveis x e y. Neste caso, definimos a variável $X = \log x$, a equação se torna $y = \alpha + \beta X$. Temos assim, um modelo de regressão linear com a variável explicada y e a variável explicativa $X = \log x$ [Mad01]. Para os dados representados pelos pontos expressos na Figura 4.3 (b), onde y cresce mais rapidamente do que x, logo, uma forma funcional possível é $y = Ae^{\beta x}$. Neste caso, podemos aplicar *log* à duas variáveis e obter outra especificação *semilog* [Mad01]:

$$\log y = \log A + \beta x$$

Se definirmos $Y = \log y$ e $\alpha = \log A$, temos

$$Y = \alpha + \beta x$$

que está na forma de uma equação de regressão linear.

Um modelo alternativo que pode ser usado é

$$y = Ax^{\beta}$$

Neste caso, aplicando *log* nos dois lados da equação, obtemos

$$\log y = \log A + \beta \log x$$

Aqui, β pode ser interpretado como uma elasticidade. Definindo $Y = \log y$, $X = \log x$ e $\alpha = \log A$, temos

$$Y = \alpha + \beta X$$

que está na forma de uma equação de regressão linear [Mad01].

Há uma diferença entre as formas funcionais nas quais transformamos a variável x e aquelas nas quais transformamos a variável y. Isto é facilmente demonstrado quando introduzimos o termo de erro u [Mad01]. Por exemplo, ao escrevermos a equação transformada com um erro aditivo, o que fazemos antes de usar o método dos mínimos quadrados, isto é, quando escrevemos

$$Y = \alpha + \beta X + u$$

estamos assumindo que a equação original, em termos de variáveis não transformadas, é

$$y = Ax^{\beta}e^{u}$$

isto é, o termo de erro entra exponencialmente e de modo multiplicativo. Se estabelecermos o pressuposto de um termo de erro aditivo na equação original, isto é

$$y = Ax^{\beta} + u$$

então não há como transformar as variáveis que nos permitiriam utilizar o método simples de estimativa aqui descrito. A estimativa deste modelo requer mínimos quadrados não-lineares [Mad01].

Pode-se lidar com não-linearidades através do que é conhecido como "procedi-

mento de busca". Suponha, por exemplo, que tenhamos a equação de regressão

$$y = \alpha + \frac{\beta}{x + \gamma} + u$$

As estimativas de α, β e γ são obtidas a partir da minimização de

$$\sum (y_i - \alpha - \frac{\beta}{x_i + \gamma})^2$$

Isto pode ser reduzido a um problema de mínimos quadrados simples da seguinte forma: para cada valor de γ, definimos a variável $Z_i = 1/(x_i + y)$ e estimamos α e β a partir da minimização de

$$\sum (y_i - \alpha - \beta z_i)^2$$

Observamos a soma dos quadrados dos resíduos em cada caso e, então, escolhemos o valor de γ para o qual a soma dos quadrados dos resíduos é mínima. As estimativas correspondentes de α e β são as estimativas dos mínimos quadrados destes parâmetros [Mad01].

Neste ponto, podemos notar como alguns problemas que, à primeira vista, não aparentam estar na estrutura de regressão simples, podem ser transformados pela redefinição das variáveis a fim de se ajustarem à referida estrutura.

4.1.2 Regressão múltipla

Em regressão simples, vimos a relação entre uma variável explicada y e uma variável explicativa x. Em regressão múltipla, veremos a relação entre y e um número de variáveis explicativas $x_1, x_2, ..., x_k$ [Mad01]. O modelo que assumimos é

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i \qquad i = 1, 2, ..., n$$

Os erros u_i são novamente conseqüência de erros de medição em y e erros es especificação da relação entre y e x. Estabelecemos as mesmas hipóteses sobre u_i. São elas

1. *Média zero.* $E(u_i) = 0$ para todo i.

2. *Variância comum.* $var(u_i) = \sigma^2$ para todo i.

3. *Independência.* u_i e u_j são independentes para todo $i \neq j$.

4. *Independência de* x_j. u_i e x_j são independentes para todo i e j.

5. *Normalidade.* u_i é normalmente distribuído para todo i.

6. Não há dependência linear entre as variáveis explicativas.

Além disto, será assumido que y_i é uma variável contínua [Mad01].

Sob as quatro primeiras hipóteses, podemos mostrar que o método dos mínimos quadrados nos dá estimadores α, β_1, β_2, ..., β_k, que são não-viesados e têm variância mínima entre a classe dos estimadores lineares não-viesados. A hipótese 5 é necessária para testes de significância e intervalos de confiança [Mad01]. Além destas hipóteses, similares as estabelecidas no caso da regressão simples, também assumiremos a hipótese 6, de que $x_1, x_2, ..., x_k$ não são colineares, isto é, não há relação linear determinística entre eles [Mad01].

Suponha, por exemplo, que tenhamos a equação de regressão

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$$

mas x_1 e x_2 estão ligados pela relação linear determinística

$$2x_1 + x_2 = 4$$

então podemos expressar x_2 em termos de x_1 e encontrar $x_2 = 4 - 2x_1$ e a equação de regressão torna-se

$$\begin{aligned} y &= \alpha + \beta_1 x_1 + \beta_2(4 - 2X_1) + u \\ &= (\alpha + 4\beta_2) + (\beta_1 - 2\beta_2)x_1 + u \end{aligned}$$

Portanto, podemos estimar $(\alpha+4\beta_2)$ e $(\beta-2\beta_2)$, mas não α, β_1 e β_2 separadamente [Mad01].

Inicialmente iremos analisar o caso de duas variáveis explicativas e, em seguida, apresentaremos as fórmulas para o caso de k variáveis explicativas.

4.1.2.1 Um modelo com duas variáveis explicativas

Considere o modelo

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \qquad i = 1, 2, ..., n \tag{4.3}$$

A hipótese que estabelecemos sobre o termo de erro u implica que

$$E(u) = 0 \qquad cov(x_1, u) = 0 \qquad cov(x_2, u) = 0$$

Como no caso do modelo de regressão simples discutido anteriormente, podemos substituir estas hipóteses por seus correspondentes amostrais [Mad01]. Sejam $\hat{\alpha}$, $\hat{\beta}_1$, e $\hat{\beta}_2$ os estimadores de α, β_1 e β_2 respectivamente, o correspondente amostral de u_i é o resíduo

$$\hat{u} = y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}$$

As três equações para se determinar α, $\hat{\beta}_1$ e $\hat{\beta}_2$ são obtidas substituindo-se as três hipóteses populacionais por seus dois correspondentes amostrais [Mad01]:

Hipóteses Populacionais	Correspondentes Amostrais
$E(u) = 0$	$(1/\text{n}) \sum \hat{u}_i = 0$ ou $\sum \hat{u}_i = 0$
$cov(u, x_1) = 0$	$(1/n) \sum x_{1i}\hat{u}_i = 0$ ou $\sum x_{1i}\hat{u}_i = 0$
$cov(u, x_2) = 0$	$(1/n) \sum x_{2i}\hat{u}_i = 0$ ou $\sum x_{2i}\hat{u}_i = 0$

Estas equações também podem ser obtidas pelo uso do método dos mínimos quadrados e são referidas como "equações normais" [Mad01].

O método dos mínimos quadrados diz que devemos escolher os estimadores de $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$ de α, β_1, β_2 de forma a minimizar

$$Q = \sum (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2$$

Diferenciando Q em relação a $\hat{\alpha}$, β_1 e β_2, igualando as derivadas a zero, encontramos

$$\frac{\partial Q}{\partial \hat{\alpha}} = 0 \Rightarrow \sum 2(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})(-1) = 0 \tag{4.4}$$

$$\frac{\partial Q}{\partial \hat{\beta}_1} = 0 \Rightarrow \sum 2(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})(-x_{1i}) = 0 \tag{4.5}$$

$$\frac{\partial Q}{\partial \hat{\beta}_2} = 0 \Rightarrow \sum 2(y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})(-x_{2i}) = 0 \tag{4.6}$$

Estas três equações, conforme mencionado anteriormente, são chamadas de "equações normais" [Mad01]. Elas podem ser simplificadas da seguinte forma:

A Equação 4.4 pode ser escrita como

$$\sum y_i = n\hat{\alpha} + \hat{\beta}_1 \sum x_{1i} + \hat{\beta}_2 \sum x_{2i}$$

ou

$$\overline{y} = \hat{\alpha} + \hat{\beta}_1 \overline{x}_1 + \hat{\beta}_2 \overline{x}_2 \tag{4.7}$$

onde

$$\overline{y} = \frac{1}{n} \sum y_i \qquad \overline{x}_1 = \frac{1}{n} \sum x_{1i} \qquad \overline{x}_2 = \frac{1}{n} \sum x_{2i}$$

A Equação 4.5 pode ser escrita como

$$\sum x_{1i} y_i = \hat{\alpha} \sum x_{1i} + \hat{\beta}_1 \sum x_{1i}^2 + \hat{\beta}_2 \sum x_{1i} x_{2i}$$

Substituindo o valor de $\hat{\alpha}$ de (4.7), temos

$$\sum x_{1i} y_i = n\overline{x}_1(\overline{y} - \hat{\beta}_1 \overline{x}_1 - \hat{beta}_2 \overline{x}_2) + \hat{\beta}_1 \sum x_{1i}^2 + \hat{\beta}_2 \sum x_{1i} x_{2i} \tag{4.8}$$

Podemos simplificar esta equação com o uso da seguinte notação. Definamos

$$
\begin{aligned}
S_{11} &= \sum x_{1i}^2 - n\overline{x}_1^2 \\
S_{12} &= \sum x_{1i}x_{2i} - n\overline{x}_1\overline{x}_2 \\
S_{22} &= \sum x_{2i}^2 - n\overline{x}_2^2 \\
S_{1y} &= \sum x_{1i}y_i - n\overline{x}_1\overline{y} \\
S_{2y} &= \sum x_{2i}y_i - n\overline{x}_2\overline{y} \\
S_{yy} &= \sum y_i^2 - n\overline{y}^2
\end{aligned}
$$

A Equação 4.8 pode ser escrita como

$$S_{1y} = \hat{\beta}_1 S_{11} + \hat{\beta}_2 S_{12} \tag{4.9}$$

Por meio de uma simplificação similar, a Equação 4.6 pode ser escrita da seguinte forma

$$S_{2y} = \hat{\beta}_1 S_{12} + \hat{\beta}_2 S_{22} \tag{4.10}$$

Agora podemos resolver estas duas equações para encontrar $\hat{\beta}_1$ e $\hat{\beta}_2$ [Mad01]. Obtemos

$$\hat{\beta}_1 = \frac{S_{22}S_{1y} - S_{12}S_{2y}}{\Delta} \tag{4.11}$$

$$\hat{\beta}_2 = \frac{S_{11}S_{2y} - S_{12}S_{1y}}{\Delta} \tag{4.12}$$

onde $\Delta = S_{11}S_{22} - S_{12}^2$. Uma vez obtidos $\hat{\beta}_1$ e $\hat{\beta}_2$, podemos obter $\hat{\alpha}$ pela Equação 4.7 [Mad01]. Assim temos

$$\widehat{\alpha} = \overline{y} - \hat{\beta}_1\overline{x}_1 - \hat{\beta}_2\overline{x}_2$$

Portanto, o procedimento computacional é o seguinte:

1. Obtenha todas as médias: $\overline{y}$, $\overline{x}_1$ e $\overline{x}_2$.

2. Obtenha todas as somas dos quadrados e dos produtos. $\sum x_{1i}^2$, $\sum x_{2i}^2$, $\sum x_{1i}x_{2i}$ e

assim por diante.

3. Obtenha S_{11}, S_{12}, S_{22}, S_{1y}, S_{2y}, S_{yy}.

4. Resolva as Equações 4.9 e 4.10 para obter $\hat{\beta}_1$ e $\hat{\beta}_2$.

5. Faça a substituição destas na Equação 4.7 para obter $\hat{\alpha}$.

Em regressão simples, definimos o seguinte [Mad01]:

$$\begin{aligned} SQR &= S_{yy} - \hat{\beta} S_{xy} \\ SQE &= \hat{\beta} S_{xy} \\ r^2_{xy} &= \frac{\hat{\beta} S_{xy}}{S_{yy}} \end{aligned}$$

As expressões análogas em regressão múltipla são

$$\begin{aligned} SQR &= S_{yy} - \hat{\beta}_1 S_{1y} - \hat{\beta}_2 S_{2y} \\ SQE &= \hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y} \\ r^2_{y\cdot 12} &= \frac{\hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y}}{S_{yy}} \end{aligned}$$

$r^2_{y\cdot 12}$ é chamado de *coeficiente de determinação múltipla* e sua raiz quadrada positiva é chamada de *coeficiente de correlação múltipla* [Mad01].

O procedimento no caso de três variáveis explicativas é análogo. As equações normais nos dão

$$\hat{\alpha} = \overline{y} - \hat{\beta}_1 \overline{x}_1 - \hat{\beta}_2 \overline{x}_2 - \hat{\beta}_3 \overline{x}_3$$

e

$$\begin{aligned} S_{1y} &= \hat{\beta}_1 S_{11} + \hat{\beta}_2 S_{12} + \hat{\beta}_3 S_{13} \\ S_{2y} &= \hat{\beta}_1 S_{12} + \hat{\beta}_2 S_{22} + \hat{\beta}_3 S_{23} \\ S_{3y} &= \hat{\beta}_1 S_{13} + \hat{\beta}_2 S_{23} + \hat{\beta}_3 S_{33} \end{aligned}$$

Novamente,

$$SQR = S_{yy} - \hat{\beta}_1 S_{1y} - \hat{\beta}_2 S_{2y} - \hat{\beta}_3 S_{3y}$$

e

$$r^2_{y \cdot 123} = \frac{\hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y} + \hat{\beta}_3 S_{3y}}{S_{yy}}$$

Note que $SQR = S_{yy}(1 - R^2)$ em todos os casos [Mad01].

Outro fator relevante mencionado anteriormente é o resíduo $\hat{u}_i$ onde temos

$$\hat{u}_i = y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}$$

Logo, as equações normais implicam que

$$\sum \hat{u}_i = 0 \qquad \sum \hat{u}_i x_{1i} = 0 \qquad \sum \hat{u}_i x_{2i} = 0 \tag{4.13}$$

Estas equações implicam que $cov(\hat{u}, x_1) = 0$ e $cov(\hat{u}, x_2) = 0$. Assim a soma dos resíduos é igual a zero e os resíduos são não-correlacionados tanto com x_1 quanto com x_2 [Mad01].

Exemplo Ilustrativo

Conforme o exemplo apresentado por G. S. Maddala [Mad01], onde o mesmo apresenta dados sobre uma amostra de cinco pessoas escolhidas ao acaso em uma grande empresa, são fornecidos dados sobre seus salários anuais, anos de educação e anos de experiência na empresa onde trabalham, conforme a Tabela 4.8.

Y	X_1	X_2	$Y - \overline{Y}$	$X_1 - \overline{X_1}$	$X_2 - \overline{X_2}$
30	4	10	0	-1	0
20	3	8	-10	-2	-2
36	6	11	6	1	1
24	4	9	-6	-1	-1
40	8	12	10	3	2

Tabela 4.8: Dados sobre salários, anos de educação e experiência [Mad01]

Y = salário anual (milhares de dólares) X_1 = anos de educação após ensino médio X_2 = anos de experiência na empresa

As médias são $\overline{Y} = 30$, $\overline{X}_1 = 5$ e $\overline{X}_2 = 10$. As somas dos quadrados dos desvios das respectivas médias são

$$S_{11} = 16$$
$$S_{12} = 12$$
$$S_{22} = 10$$
$$S_{1y} = 62$$
$$S_{2y} = 52$$
$$S_{yy} = 272$$

As equações normais são

$$16\hat{\beta}_1 + 12\hat{\beta}_2 = 62$$
$$12\hat{\beta}_1 + 10\hat{\beta}_2 = 52$$

Resolvendo estas equações temos

$$\hat{\beta}_1 = -0,25$$
$$\hat{\beta}_2 = 5,5$$
$$\hat{\alpha} = \overline{Y} - \hat{\beta}_1\overline{X}_1 - \hat{\beta}_2\overline{X}_2 = 30 - (-1,25) - 55 = -23,75$$
$$R^2 = \frac{\hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y}}{S_{yy}} = \frac{271,5}{272} = 0,998$$

Logo, a equação de regressão é

$$\hat{Y} = -23,75 - 0,25X_1 + 5,5X_2 \qquad R^2 = 0,998$$

Fica claro neste exemplo, que estes números contém alguma incerteza, devido

principalmente ao multiplicador negativo $-0,25$ junto à variável X_1, que representa os anos de educação após o ensino médio, significando que quanto mais anos de educação após o ensino médio certa pessoa possuir, menor será seu salário [Mad01]. O termo constante, $-23,75$ também pode pode parecer estranho, pois indicaria um salário negativo para um novo funcionário. Neste caso, devemos lembrar que não devemos extrapolar os resultados muito além do intervalo amostral, já que nossa amostra não inclui novos funcionários [Mad01].

O que podemos concluir é que a amostra que analisamos não representa todas as pessoas que trabalham na empresa, e deve ter sido tirada de um sub-grupo específico [Mad01].

4.1.2.2 Inferência estatística no modelo de regressão múltipla

Considerando inicialmente os resultados de um modelo com duas variáveis explicativas, e assumindo que os erros u_i são normalmente distribuídos, isto, juntamente com os outros pressupostos estabelecidos, implica que u_i são independentes, normalmente distribuídos com média zero e possuem uma variância comum σ^2 [Mad01]. Baseado nestes pressupostos, os seguintes resultados podem ser derivados:

1. $\hat{\alpha}$, $\hat{\beta}_1$ e $\hat{\beta}_2$ têm distribuições normais com médias α, β_1, β_2, respectivamente
2. Se denotarmos o coeficiente de correlação entre x_1 e x_2 por r_{12}, então

$$
\begin{aligned}
var(\hat{\beta}_1) &= \frac{\sigma^2}{S_{11}(1-r_{12}^2)} \\
var(\hat{\beta}_2) &= \frac{\sigma^2}{S_{22}(1-r_{12}^2)} \\
cov(\hat{\beta}_1, \hat{\beta}_2) &= \frac{-\sigma^2 r_{12}^2}{S_{12}(1-r_{12}^2)} \\
var(\hat{\alpha}) &= \frac{\sigma^2}{n} + \overline{x}_1^2 var(\hat{\beta}_1) + 2\overline{x}_1\overline{x}_2 cov(\hat{\beta}_1, \hat{\beta}_2) + \overline{x}_2^2 var(\hat{\beta}_2) \\
cov(\hat{\alpha}, \hat{\beta}_1) &= -[\overline{x}_1 var(\hat{\beta}_1) + \overline{x}_2 cov(\hat{\beta}_1, \hat{\beta}_2)] \\
cov(\hat{\alpha}, \hat{\beta}_2) &= -[\overline{x}_1 cov(\hat{\beta}_1, \hat{\beta}_2) + \overline{x}_2 var(\hat{\beta}_2)]
\end{aligned}
$$

Analogamente aos outros resultados no caso da regressão simples, temos os seguintes resultados:

3. Se SQR for a soma dos quadrados dos resíduos, então SQR/σ^2 tem uma distribuição chi-quadrado com $(n-3)$ graus de liberdade. Este resultado pode ser utilizado ao fazer afirmações sobre os intervalos de confiança sobre σ^2.

4. Se $\hat{\sigma}^2 = SQR/(n-3)$, então $E(\hat{\sigma}^2) = \sigma^2$ ou, $\hat{\sigma}^2$ é um estimador não-viesado de σ^2 [Mad01].

5. Se substituirmos $\hat{\sigma}^2$ por σ^2 nas expressões do resultado 2, teremos as variâncias e covariâncias estimadas. As raizes quadradas das variâncias estimadas são chamadas de erros padrão (EP), então

$$\frac{\hat{\alpha}-\alpha}{EP(\hat{\alpha})} \qquad \frac{\hat{\beta}_1-\beta_1}{EP(\hat{\beta}_1)} \qquad \frac{\hat{\beta}_2-\beta_2}{EP(\hat{\beta}_2)}$$

   cada um tem uma distribuição t com $(n-3)$ graus de liberdade [Mad01].

Além dos resultados que têm correspondentes no caso de regressão simples, há um ítem extra no caso de regressão múltipla, que se refere às regiões de confiança e aos testes conjuntos de parâmetros. Temos o seguinte resultado:

6. $F = (1/2\sigma^2)[S_{11}(\hat{\beta}_1-\beta_1)^2 + 2S_{12}(\hat{\beta}_1-\beta_1)(\hat{\beta}_2-\beta_2) + S_{22}(\hat{\beta}_2-\beta_2)^2]$ tem uma distribuição F com 2 e $(n-3)$ graus de liberdade. Este resultado pode ser utilizado para se construir uma região de confiança para β_1 e β_2 juntos e para testar β_1 e β_2 conjuntamente [Mad01].

Exemplo Ilustrativo

Utilizaremos o exemplo citado por G. S. Maddala [Mad01] para demonstrar o procedimento. Uma função de produção é especificada como

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

onde

$$y = \text{log produção}$$
$$x_1 = \text{log trabalho utilizado}$$
$$x_2 = \text{log capital utilizado}$$

As variáveis x_i são não-estocásticas. Os seguintes dados são obtidos de uma amostra de tamanho $n = 23$:

$$\begin{array}{lll} \overline{x}_1 = 10 & \overline{x}_2 = 5 & \overline{y} = 12 \\ S_{11} = 12 & S_{12} = 8 & S_{22} = 12 \\ S_{1y} = 10 & S_{2y} = 8 & S_{yy} = 10 \end{array}$$

As equações normais são

$$\begin{aligned} 12\hat{\beta}_2 + 8\hat{\beta}_2 &= 10 \\ 8\hat{\beta}_2 + 12\hat{\beta}_2 &= 8 \end{aligned}$$

Isto nos dá $\hat{\beta}_1 = 0,7$ e $\hat{\beta}_2 = 0,2$. Logo,

$$\begin{aligned} \hat{\alpha} &= \overline{y} - \hat{\beta}_1\overline{x}_1 - \hat{\beta}_2\overline{x}_2 = 12 - 0,7(10) - 0,2(5) = 4 \\ R^2_{y\cdot 12} &= \frac{\hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y}}{S_{yy}} \\ &= \frac{0,7(10) + 0,2(8)}{10} = 0,86 \\ SQR &= S_{yy}(1 - R^2) = 10(1 - 0,86) = 1,4 \end{aligned}$$

Logo

$$\begin{aligned} \sigma^2 &= \frac{SQR}{n-3} = \frac{1,4}{20} = 0,07 \\ r^2_{12} &= \frac{S^2_{12}}{S_{11}S_{22}} = \frac{64}{144} \end{aligned}$$

Então temos

$$S_{11}(1 - r^2_{12}) = 12(\frac{80}{144}) = \frac{80}{12} = \frac{20}{3}$$

Logo

$$var(\hat{\beta}_1) = \frac{\sigma^2}{S_{11}(1 - r_{12}^2)} = \frac{3}{20}\sigma^2$$
$$var(\hat{\beta}_2) = \frac{3}{20}\sigma^2$$

e

$$cov(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\sigma^2 r_{12}^2}{S_{12}(1 - r_{12})}$$
$$= \frac{-\sigma^2(64/144)}{8(80/144)} = -\frac{\sigma^2}{10}$$

Além disto, como $\overline{x}_1 = 10$ e $\overline{2}_2 = 5$, temos

$$V(\hat{\alpha}) = \sigma^2[\frac{1}{23} + (10)^2(\frac{3}{20}) - \frac{2(10)(5)}{10} + \frac{(5)^2(3)}{20}]$$
$$= 8,7935\sigma^2$$

Substituindo a estimativa de σ^2, que é $0,07$ nestas expressões e extraindo as raízes quadradas, temos

$$EP(\hat{\beta}_1) = EP(\hat{\beta}_1) = \sqrt{\frac{0,21}{20}} = 0,102$$
$$EP(\hat{\alpha}) = 0,78$$

Logo, a equação de regressão é

$$\hat{y} = 4,0 + 0,7x_1 + 0,2x_2 \qquad R^2 = 0,86$$

Utilizando a distrubuição t, com 20 graus de liberdade, temos os intervalos de confiança de 95% para α, β_1 e β_2 como

$$\hat{\alpha} \pm 2,086EP(\hat{\alpha}) = 4,0 \pm 1,63 = (2,37; 5,63)$$
$$\hat{\beta_1} \pm 2,086EP(\hat{\beta_1}) = 0,7 \pm 0,21 = (0,49; 0,91)$$
$$\hat{\beta_2} \pm 2,086EP(\hat{\beta_2}) = 0,2 \pm 0,21 = (-0,01; 0,41)$$

Fórmulas para o caso geral de K variáveis explicativas

Apresentamos anteriormente exemplos de expressões explícitas par ao caso de duas variáveis explicativas de forma a destacar as diferenças entre regressão simples e múltipla. Se tivermos a equação de regressão múltipla com k regressores, isto é

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

então temos que implementar as seguintes mudanças nos resultados anteriores. No resultado 2, nas expressões para $V(\hat{\beta}_1)$, $V(\hat{\beta}_2)$ e assim por diante, o denominador será a soma doas quadrados dos resíduos da regressão da variável em todos os outros regressores [Mad01]. Logo

$$var(\hat{\beta}_i) = \frac{\sigma^2}{SQR_1} \text{ para } i = 1, 2, ..., k$$

onde SQR_i é uma soma dos quadrados dos resíduos de uma regressão de x_i em todos outros $(k-1)$ regressores. Estas k regressões são chamadas regressões auxiliares [Mad01].

Nos resultados 3, 4 e 5, temos que mudar os graus de liberdade de $(n-3)$ para $(n-k-1)$. Em regressão simples, temos $k = 1$ e, portanto $(n-2)$ graus de liberdade. No caso de duas variáveis explicativas, $k = 2$ e, por conseguinte $(n-3)$ graus de liberdade. Em todos os casos, um grau de liberdade é para a estimativa de α [Mad01].

O resultado 6 agora passa a ser

$$f = \frac{1}{k\hat{\sigma}^2}[\sum\sum S_{ij}(\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j)]$$

que tem uma distribuição F com k e $(n-k-1)$ graus de liberdade [Mad01].

Exemplo Ilustrativo

Conforme o exemplo apresentado por G. S. Maddala [Mad01], em junho de 1978, os eleitores da Califórnia aprovaram o que é conhecido como Proposição 13, limitando os impostos de propriedade. Isto levou a reduções substanciais e diferenciadas nos impostos de propriedade, o que teve como conseqüência o aumento dos preços imobiliários. Kenneth T. Rosen [Ros82] estudou o impacto da redução dos impostos de propriedade nos preços imobiliários na área da baía de São Francisco. Além de impostos de propriedade, existem outros fatores que determinam os preços imobiliários e eles devem ser considerados no estudo. sendo assim, Rosem incluiu características do imóvel, tais como quantidade de metros quadrados, a idade da casa e um índice de qualidade. Ele também incluiu fatores econômicos, tais como renda média e tempo de transporte para São Francisco. As equações estimadas foram

$$\hat{y} = o, 171 + 7, 275x_1 + 0, 547x_2 + 0, 00073x_3 + 0, 0638x_4 - 0, 0043x_5 + 0, 857x_6$$
$$R^2 = 0, 897$$
$$n = 64$$

onde

$y =$ Mudança na média dos preços imobiliários pós-Proposição 13
$x_1 =$ Redução da tributação pós-Proposição 13
$x_2 =$ Tamanho médio do imóvel
$x_3 =$ Renda média das famílias na vizinhança
$x_4 =$ Idade média do imóvel
$x_5 =$ Tempo de transporte para São Francisco
$x_6 =$ Índice de qualidade do imóvel conforme avaliado por corretores

Todos os coeficientes têm os sinais esperados.

O coeficiente x_1 indica que cada diminuição de de \$1 nos impostos de propriedade, aumento os valores da propriedade em \$7. A questão é se isso é próximo a magnitude certa. Assumindo que se espera que as reduções dos impostos de propriedade estejam no mesmo nível nos anos futuros, o valor presente de um retorno de \$1 por ano é $1/r$, onde r é a taxa de juros. Isso é igual a \$7 se $r = 14, 29\%$. A taxa de juros estava na época ao

redor deste nível, e então, Rosen concluiu: "A taxa de capitalização determinada por essa equação é cerca de 7%. exatamente a magnitude que se deve esperar com uma taxa de juros de 12 a 15%."

4.1.2.3 Interpretação dos coeficientes de regressão

Em regressão simples estamos interessados em medir os efeitos da variável explicativa na variável explicada [Mad01]. Já que a equação de regressão pode ser escrita como

$$y - \overline{y} = \hat{\beta}(x - \overline{x}) + \hat{u}_i$$

esse efeito é medido por $\hat{\beta}$, onde

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \text{ ou } \frac{cov(x,y)}{V(x)}$$

Na equação de regressão múltipla com duas variáveis explicativas, x_1 e x_2, podemos falar do efeito conjunto de x_1 e x_2 e do efeito parcial de x_1 ou x_2 sobre y [Mad01]. Como a equação de regressão pode ser escrita como

$$y - \overline{y} = \hat{\beta}_1(x_1 - \overline{x_1}) + \hat{\beta}_2(x_2 - \overline{x_2}) + \hat{u}_i$$

o efeito parcial de x_1 é medido por $\hat{\beta}_1$ e o efeito parcial de x_2 é medido por $\hat{\beta}_2$. Por efeito parcial queremos dizer que mantemos as outras variáveis constantes ou que eliminamos seus efeitos[Mad01]. Logo $\hat{\beta}_1$ deve ser interpretado como a medição do efeito de x_1 em y após se eliminarem os efeitos de x_2 em x_1. Similarmente $\hat{\beta}_2$ deve ser interpretado como a medição do efeito de x_2 em y após se eliminarem os efeitos de x_1 em x_2. Essa interpretação sugere que podemos derivar o estimador $\hat{\beta}_1$ de β_1 estimando suas regressões simples separadas [Mad01]:

1. Estime uma regressão de x_1 em x_2. Denote o coeficiente de regressão por β_{12}. Denotando os resíduos dessa equação por W_i, temos

$$W_i = x_{1i} - \overline{x}_1 - b_{12}(x_{2i} - \overline{x}_2) \tag{4.14}$$

Note que W_i é a parte de x_i restante após a remoção do efeito de x_2 em x_1.

2. Agora regrida y_i em W_i. O coeficiente de regressão é justamente β_1, o qual derivamos antes da regressão múltipla.

Suponha que também eliminemos o efeito de x_2 em y. Seja V_i o resíduo da regressão de y em x_2. Se agora regredirmos V_i em W_i, então o coeficiente de regressão resultante será o mesmo que o obtido da regressão de y_i em W_i [Mad01]. Isso ocorre porque

$$V_i = y_i - \overline{y} - b_{y2}(x_{2i} - \overline{x_2}$$

onde b_{y2} é o coeficiente de regressão de y em x_2. Entretanto, x_2 será não-correlacionado com W [Mad01]. Então

$$cov(V, W) = cov(y, W)$$

Logo, a regressão de V_i em W_i produzirá o mesmo estimador $\hat{\beta}_1$ como uma regressão de y_i em W_i [Mad01].

Esse resultado é importante e útil na eliminação de tendência e em ajusto sazonal de dados temporais [Mad01]. Isso implica que se tivermos uma variável explicada y e uma variável explicativa x e existir outra variável Z inconveniente que influencia tanto y quanto x, o efeito "puro" de x em y, após se eliminar o efeito desta variável Z inconveniente em x e y, poderá ser calculado simplesmente através da estimativa da equação de regressão múltipla [Mad01].

$$y = \alpha + \beta x + \gamma Z + u$$

O coeficiente β nos dá o efeito necessário "puro". Não temos que regredir y em Z e x em Z para eliminar o efeito de Z nessas variáveis e, então, fazer uma terceira regressão de y em x. Embora tenhamos demonstrado o resultado para apenas duas variáveis, ele é generalizável. Tanto x quanto Z podem ser conjuntos de variáveis [Mad01].

4.1.2.4 Previsão no modelo de regressão múltipla

As formulas para previsão em regressão múltipla são similares as formulas apresentadas no caso de regressão simples, exceto pelo fato de que para se calcular o erro padrão dos valores previsto precisamos das variâncias e covariâncias de todos os coefici-

entes de regressão [Mad01]. Novamente, apresentamos a expressão para o erro padrão no caso de duas variáveis explicativas e então a expressão para o caso geral de k variáveis explicativas. Seja a equação de regressão estimada igual a

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Agora considere a previsão do valor y_0 de y, dados os valores x_{10} de x_1, e x_{20} de x_2, respectivamente.

Temos então

$$y_0 = \alpha + \beta_1 x_{10} + \beta_2 x_{20} + u_0$$

Considere

$$\hat{y}_0 = \hat{\alpha} + \hat{\beta}_1 x_{10} + \hat{\beta}_2 x_{20}$$

O erro de previsão é

$$\hat{y}_0 - y_0 = \hat{\alpha} - \alpha + (\hat{\beta}_1 - \beta_1)x_{10} + (\hat{\beta}_2 - \beta_2)x_{20} - u_0$$

Como $E(\hat{\alpha} - \alpha)$, $E(\hat{\beta}_1 - \beta_1)$, $E(\hat{\beta}_2 - \beta_2)$ e $E(u_0)$ são todos iguais a zero, temos $E(\hat{y}_0 - y_0) = 0$. Logo o previsor $\hat{y}_0$ é não-viesado. Note que estamos dizendo que $E(y_0) = E(\hat{y}_0)$ [Mad01]. A variância do erro de previsão é

$$\begin{aligned}\sigma^2(1 + \frac{1}{n}) &+ (x_{10} - \overline{x}_1)^2 var(\hat{\beta}_1) \\ &+ 2(x_{10} - \overline{x}_1)(x_{20} - \overline{x}_2) cov(\hat{\beta}_1, \hat{\beta}_2) \\ &+ (x_{20} - \overline{x}_2)^2 var(\hat{\beta}_2)\end{aligned}$$

No caso de k variáveis explicativas,

$$\sigma^2(1 + \frac{1}{n}) + \sum\sum (x_{i0} - \overline{x}_i)(x_{j0} - \overline{x}_j) cov(\hat{\beta}_i, \hat{\beta}_j)$$

Estimamos σ^2 por $SQR/(n-3)$ no caso de duas variáveis explicativas e por

$SQR/(n-k-1)$ no caso geral [Mad01].

Exemplo Ilustrativo

Considere o exemplo citado anteriormente, onde a reta de regressão é

$$y = 4,0 + 0,7x_1 + 0,2x_2$$

Considere a previsão de y para $x_{10} = 12$, e $x_{20} = 7$. Temos

$$\hat{y}_0 = 4,0 + 0,7(12) + 0,2(7) = 13,8$$

Note que

$$x_{10} - \overline{x}_1 = 12 - 10 = 2$$
$$x_{20} - \overline{x}_2 = 7 - 5 = 2$$

Usando as expressões para $var(\hat{\beta}_1)$, $var(\hat{\beta}_2)$, $cov(\hat{\beta}_1, \hat{\beta}_2)$ e σ^2, temos a variância estimada do erro de previsão como

$$0,07(1 + \frac{1}{23}) + 4(\frac{3}{20} + \frac{3}{20} - \frac{2}{10})(0,07) = 0,101$$

O desvio padrão da previsão é 0,318. Logo, o intervalo de confiança de 95% para a previsão é

$$13,8 \pm 2,086(0,318) \qquad \text{ou} \qquad 13,8 \pm 0,66 \qquad \text{ou} \qquad (13,14; 14,46)$$

4.1.2.5 Omissão de variáveis relevantes e inclusão de variáveis irrelevantes

Até agora assumimos que a equação de regressão múltipla que estamos estimando inclui todas as variáveis explicativas relevantes. Na prática, esse raramente é o caso. Às vezes, algumas variáveis relevantes não são incluídas devido a equívocos ou falhas de medição. Em outros momentos, algumas variáveis irrelevantes são incluídas [Mad01]. O que gostaríamos de saber é como nossas inferências mudam quando esses problemas estão presentes.

Omissão de variáveis relevantes

Consideremos primeiro a omissão de variáveis relevantes. Suponha que a equação verdadeira seja

$$y = \beta_1 x_1 + \beta_2 x_2 + u \tag{4.15}$$

Em vez disso, omitimos x_2 e estimamos a equação

$$y = \beta_1 x_1 + e$$

Isso será referido como o "modelo erroneamente mal especificado" [Mad01]. O estimador de β_1 que temos é

$$\hat{\beta}_1 = \frac{\sum x_1 y}{\sum x_1^2}$$

Substituindo aqui a expressão por y da Equação 4.1.2.5, temos

$$\hat{\beta}_1 = \frac{\sum x_1(\beta_1 x_1 + \beta_2 x_2 + u)}{\sum x_1^2} = \beta_1 + \beta_2 \frac{\sum x_1 x_2}{\sum x_1^2} + \frac{\sum x_1 u}{\sum x_1^2}$$

Como $E(\sum x_1 u) = 0$, temos

$$E(\hat{\beta}_1) = \beta_1 + b_{21}\beta_2 \tag{4.16}$$

onde $b_{21} = \sum x_1 x_2 / \sum x_1^2$ é o coeficiente da regressão de x_2 em x_1.

Logo, $\hat{\beta}_1$ é um estimador viesado de β_1 e o viés é dado por

$$\text{desvio} = \text{coeficiente da variável excluída} * \text{coeficiente de regressão em uma regressão variável}$$

Se denotarmos o estimador de β_1 da Equação 4.1.2.5 por $\tilde{\beta}_1$, a variância de $\tilde{\beta}_1$ é dada por

$$var(\tilde{\beta}_1) = \frac{\sigma^2}{S_1 1(1 - r_{12}^2)}$$

onde

$$S_1 1 = \sum x_1^2$$

Por outro lado,

$$var(\hat{\beta}_1) = \frac{\sigma^2}{S_1 1}$$

Logo, $\hat{\beta}_1$ é um estimador viesado mais tem uma variância menor do que $\tilde{\beta}_1$ [Mad01]. Na verdade, a variância seria consideravelmente menor se r_{12}^2 fosse menor. No entanto, o erro padrão estimado não necessariamente será menor para $\hat{\beta}_1$ do que é para $\tilde{\beta}_1$. Isso ocorre porque σ^2, a variância estimada do erro, pode ser maior no modelo erroneamente mal especificado. Ela é dada pela soma dos quadrados dos resíduos dividida por graus de liberdade, e pode ser maior ou menor para o modelo erroneamente mal especificado [Mad01]. Denotemos a variância estimada por S^2. Então a fórmula ligando as vairÂncias estimadas é

$$\frac{S^2(\hat{\beta}_1)}{S^2(\tilde{\beta}_1)} = \frac{1 - r_{12}^2}{1 - r_{y2\dot{1}}^2}$$

Logo, o erro padrão de $\hat{\beta}_1$ será menor do que o erro padrão de $\tilde{\beta}_1$ apenas se $r_{12}^2 > r_{y2\dot{1}}^2$ [Mad01].

Consideramos o caso de apenas uma variável incluída e uma variável omitida. No caso em que temos $k-1$ variáveis incluídas e a k-ésima variável omitida, a Equação 4.16 generaliza para

$$E(\hat{\beta}_i) = \beta_i + b_{ki}\beta_k \qquad\qquad i = 1, 2, ..., k-1 \tag{4.17}$$

onde b_{ki} é o coeficiente de regressão de x_i na regressão auxiliar de x_k em x_1, x_2, ..., x_{k-1}. Isto é, consideramos a regressão da variável omitida x_k em todas as variáveis incluídas [Mad01].

No caso geral, onde temos diversas variáveis incluídas e diversas variáveis omitidas, temos que estimar as regressões múltiplas "auxiliares" de cada uma das variáveis excluídas em todas as variáveis incluídas [Mad01]. O viés em cada um dos coeficientes

estimados das variáveis incluídas será uma média ponderada dos coeficientes de todas as variáveis excluídas com pesos obtidos das regressões múltiplas auxiliares [Mad01].

Suponha que tenhamos k variáveis explicativas, das quais as primeiras k_1 são são incluídas e as $(k-k_1)$ restantes são omitidas. Então a fórmula correspondente às Equações 4.16 e 4.17 é

$$E(\hat{\beta}_i) = \beta_i + \sum b_{ji}\beta_j \qquad j = k_1 + 1 \qquad i = 1, 2, ..., k_1 \tag{4.18}$$

onde b_{ji} é o coeficiente de regressão da *i-ésima* variável incluída na regressão da *j-ésima* variável omitida em todas as variáveis incluídas. As Equações 4.16 4.18 pode ser utilizadas para gerar estimativas da direção dos vieses nos coeficientes estimados quando algumas variáveis são omitidas por causa da falha de observação ou porque elas não são mensuráveis [Mad01].

Exemplo Ilustrativo

Utilizando o exemplo citado por G. S. Maddala [Mad01], considere a estimativa da demanda por alimentos nos Estados Unidos com base na Tabela 4.9.

Onde Q_D representa o consumo de alimento *Per Capita*, Q_S representa a produção de alimento *Per Capita*, P_D representa os preços de alimento no varejo / índice do custo de vida, Y representa a renda disponível / índice de custo de vida, P_S representa os preços recebidos pelos fazendeiros por alimentos / índice do custo de vida e t representa o tempo.

Suponha que a equação "verdadeira" seja

$$Q_D = \alpha + \beta_1 P_D + \beta_2 Y + u$$

Entretanto, omitimos a variável renda. Temos

$$\hat{Q}_D = 89,97 + 0,107 P_D \qquad \hat{\sigma}^2 = 2,338$$

O coeficiente P_D tem o sinal errado. Isso pode ser atribuído à omissão da variável renda? A resposta é sim, uma vez que o coeficiente de P_D tem uma estimativa viesada com o viés dado por

Ano	Q_D	P_D	Y	Q_S	P_S	t
1922	98,6	100,2	87,4	108,5	99,1	1
1923	101,2	101,6	97,6	110,1	99,1	2
1924	102,4	100,5	96,7	110,4	98,9	3
1925	100,9	106,0	98,2	104,3	110,8	4
1926	102,3	108,7	99,8	107,2	108,2	5
1927	101,5	106,7	100,5	105,8	105,6	6
1928	101,6	106,7	103,2	107,8	109,8	7
1929	101,6	108,2	107,8	103,4	108,7	8
1930	99,8	105,5	96,6	102,7	100,6	9
1931	100,3	95,6	88,9	104,1	81,0	10
1932	97,6	88,6	75,1	99,2	68,6	11
1933	97,2	91,0	76,9	99,7	70,9	12
1934	97,3	97,9	84,6	102,0	81,4	13
1935	96,0	102,3	90,6	94,3	102,3	14
1936	99,2	102,2	103,1	97,7	105,0	15
1937	100,3	102,5	105,1	101,1	110,5	16
1938	100,3	97,0	96,4	102,3	92,5	17
1939	104,1	95,8	104,4	104,4	89,3	18
1940	105,3	96,4	110,7	108,5	93,0	19
1941	107,6	100,3	127,1	111,3	106,6	20

Tabela 4.9: Dados sobre demanda e oferta de alimentos nos Estados Unidos [Mad01]

$$\begin{aligned}\text{desvio} &= \text{coeficiente da variável renda} \\ &\quad * \text{coeficiente de regressão da renda no preço}\end{aligned}$$

Espera-se que o coeficiente de renda seja positivo. Além disso, uma vez que os dados são séries temporais, esperaríamos uma correlação positiva entre P_D e Y. Portanto, espera-se que o viés seja positivo, e isso pode transformar um coeficiente negativo em um positivo.

Nesse caso, a equação de regressão com Y incluído dá o resultado

$$\hat{Q}_D = 92,05 - 0,142P_D + 0,236Y \qquad \hat{\sigma}^2 = 1,952$$

Note que o coeficiente de P_D agora é negativo.

Inclusão de variáveis irrelevantes

Considere agora o caso da inclusão de variáveis irrelevantes. Suponha que a equação verdadeira seja

$$y = \beta_1 x_1 + u$$

mas estimamos a equação

$$y = \beta_1 x_1 + \beta_2 x_2 + v$$

Os estimadores dos mínimos quadrados $\tilde{\beta}_1$ e $\tilde{\beta}_2$ dessa equação "erroneamente" especificada são dados por

$$\tilde{\beta}_1 = \frac{S_{22}S_{1y} - S_{12}S_{2y}}{S_{11}S_{22} - S_{12}^2}$$
$$\tilde{\beta}_2 = \frac{S_{11}S_{2y} - S_{12}S_{1y}}{S_{11}S_{22} - S_{12}^2}$$

onde $S_{11} = \sum x_1^2$, $S_{1y} = \sum x_1 y$, $S_{12} = \sum x_1 x_2$, e assim por diante [Mad01].

Como $y = \beta_1 x_1 + u$, temos $E(S_{2y}) = \beta_1 S_{12}$ e $E(S_{1y}) = \beta_1 S_{11}$.

Assim temos

$$E(\tilde{\beta}_1) = \beta_1 \qquad \text{e} \qquad E(\tilde{\beta}_2) = 0$$

Logo, temos estimativas não-viesadas para ambos os parâmetros [Mad01]. Esse resultado, ligado aos resultados anteriores relacionados ao desvio introduzido pela omissão das variáveis relevantes, poderia nos levar a acreditar que é melhor incluir variáveis em vez de excluí-las. Todavia, isso não procede, pois apesar da inclusão de variáveis irrelevantes não ter efeito no viés dos estimadores, ela afeta as variâncias [Mad01].

A variância de $\hat{\beta}_1$, o estimador de β_1 da equação correta é dada por

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{11}}$$

Por outro lado, da equação erroneamente especificada temos

$$var(\tilde{\beta}_1) = \frac{\sigma^2}{(1 - r_{12}^2)S_{11}}$$

onde r_{12} é a correlação entre x_1 e x_2. Logo, $var(\tilde{\beta}_1) > var(\hat{\beta}_1)$, a não ser que $r_{12} = 0$. Conseqüentemente, acharemos estimadores não-viesados mais ineficientes ao incluir a variável irrelevante [Mad01].

4.2 Variáveis *Dummy*

Nesta sessão, analisaremos um tipo especial de variável que é utilizada em equações de regressão múltipla, as variáveis *Dummy*, e os problemas por ela causados.

As variáveis explicativas *dummy*, podem ser utilizadas para propósitos diferentes, tais como:

1. Permitir diferenças nos termos dos interceptos.
2. Permitir diferenças nos coeficientes angulares.
3. Estimar equações com restrições de equações cruzadas.
4. Testar a estabilidade dos coeficientes de regressão.

Analisaremos alguns destes objetivos a seguir.

4.2.1 Variáveis *Dummy* para mudanças nos termos do intercepto

Em alguns casos, há algumas variáveis explicativas em nossa equação de regressão que são apenas qualitativas. Em tais casos, com freqüência, levam-se em conta esses efeitos por meio de variáveis *dummy* [Mad01]. O pressuposto implícito é que as retas de regressão para os diferentes grupos diferem apenas no termo de intercepto, tendo-se os mesmos coeficientes angulares [Mad01]. Suponha, por exemplo, que a relação entre a renda y e os anos de escolaridade x para dos grupos se dá conforme mostrado na Figura 4.4. Os pontos referem-se ao grupo 1 e os círculos ao grupo 2.

Note que as inclinações das retas de regressão para ambos os grupos são aproximadamente as mesmas, mas os interceptos são diferentes. Portanto, as equações de regressão que elaboramos serão

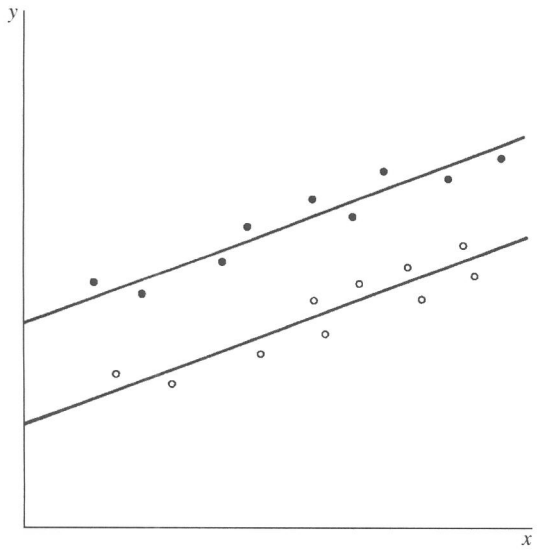


Figura 4.4: Regressão com inclinação comum e interceptos diferentes [Mad01]

$$y = \begin{cases} \alpha_1 + \beta x + u & \text{para o primeiro grupo} \\ \alpha_2 + \beta x + u & \text{para o segundo grupo} \end{cases} \tag{4.19}$$

Essas equações podem ser combinadas em uma única equação

$$y = \alpha_1 + (\alpha_2 - \alpha_1)D + \beta x + u \tag{4.20}$$

onde

$$D = \begin{cases} 1 & \text{para o grupo 2} \\ 0 & \text{para o grupo 1} \end{cases}$$

A variável D é uma variável *dummy*. O coeficiente da variável *dummy* mede as diferenças entre os dois termos do intercepto [Mad01].

Se houver mais grupos, precisaremos introduzir mais variáveis *dummy*. Para três grupos, temos

$$y = \begin{cases} \alpha_1 + \beta x + u & \text{para o o grupo 1} \\ \alpha_2 + \beta x + u & \text{para o o grupo 2} \\ \alpha_3 + \beta x + u & \text{para o o grupo 3} \end{cases}$$

Isso pode ser escrito como

$$y = \alpha_1(\alpha_2 - \alpha_1)D_1 + (\alpha_3 - \alpha_1)D_2 + \beta x + u \tag{4.21}$$

onde

$$D_1 = \begin{cases} 1 & \text{para o grupo 2} \\ 0 & \text{para o grupo 1 e 3} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{para o grupo 3} \\ 0 & \text{para o grupo 1 e 2} \end{cases}$$

Pode-se ver facilmente que ao substituir os valores de D_1 e D_2 na Equação 4.21, encontramos os interceptos α_1, α_2, α_3, respectivamente, para os três grupos. Note que ao combinar estas três equações, estamos assumindo que o coeficiente angular β é o mesmo para todos os grupos e que o termo de erro u tem a mesma distribuição para os três grupos [Mad01].

Se houver um termo constante na equação de regressão, o número de variáveis *dummy* definidas deve sempre ser um a menos do que o número de agrupamentos da categoria, porque o termo constante é o intercepto para o grupo o grupo base e os coeficientes das variáveis *dummy* medem as diferenças nos interceptos, conforme pode-se ver na Equação 4.21 [Mad01]. Nessa equação, o termo constante mede o intercepto para o primeiro grupo, o termo constante mais o coeficiente D_1 mede o intercepto do segundo grupo e o termo constante mais o coeficiente D_2 mede o intercepto do terceiro grupo. Escolhemos o grupo 1 como o grupo base, mas pode-se escolher qualquer um. Os coeficientes das variáveis *dummy* medem as diferenças entre os interceptos daquelas provenientes do grupo

base. Se não introduzirmos um termo constante na equação de regressão, podemos definir uma variável *dummy* para cada grupo, e interceptos dos respectivos grupos [Mad01].

Como outro exemplo, suponha que tenhamos séries sobre consumo C e renda Y para um número de famílias. Além disto, temos séries sobre

1. S: o sexo do chefe da família.
2. A: a idade do chefe da família, dada em três categorias, menos de 25 anos, 25 a 50 anos e mais de 50 anos.
3. E: a educação do chefe da família, também em três categorias, menor que nível médio, maior ou igual ao nível médio, mas menor que nível superior e maior do que nível superior.

Incluímos essas variáveis qualitativas na forma de variáveis *dummy*.

$$D_1 = \begin{cases} 1 & \text{se homem} \\ 0 & \text{se mulher} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{se idade} < 25 \text{ anos} \\ 0 & \text{caso contrário} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{se idade estiver entre 25 e 50 anos} \\ 0 & \text{caso contrário} \end{cases}$$

$$D_4 = \begin{cases} 1 & \text{se} < \text{ensino médio} \\ 0 & \text{caso contrário} \end{cases}$$

$$D_5 = \begin{cases} 1 & \text{se} \geq \text{ensino médio mas} < \text{superior} \\ 0 & \text{caso contrário} \end{cases}$$

Para cada categoria, o número de variáveis *dummy* é um a menos do que o número de classificações [Mad01].

Então, processamos a equação de regressão

$$C = \alpha + \beta Y + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \gamma_4 D_4 + \gamma_5 D_5 + u$$

O pressuposto estabelecido sobre o método da variável *dummy* é que apenas o intercepto muda para cada grupo, e não os coeficientes angulares [Mad01]. Obtém-se o

termo de intercepto para cada indivíduo substituindo os valores apropriados para D_1 até D_5. Por exemplo, para um homem, idade < 25, com nível superior, temos $D_1 = 1$, $D_2 = 1$, $D_3 = 0$, $D_4 = 0$, $D_5 = 0$, e, portanto, o intercepto é $\alpha + \gamma_1 + \gamma_2$.

Exemplo Ilustrativo

Utilizando o exemplo citado por G. S. Maddala [Mad01], imagine o seguinte caso. A Agência de Proteção Ambiental, do inglês (EPA, publica estimativas de milhagem desenvolvidas para ajudar compradores de carros na comparação de eficiência relativa de consumo de combustível para diferentes modelos. A estimativa da EPA provê todas as informações necessárias para se comparar a eficiência relativa de consume de combustível desses diferentes modelos? Para investiga este problema, Lovell [Lov86] estimou as seguintes regressões:

$$\hat{y} = 7,952 + 0,693EPA \qquad \overline{R}^2 = 0,74$$

$$\hat{y} = 22,008 - 0,002W - 2,760S/A + 2,280G/D + 0,415EPA \qquad \overline{R}^2 = 0,83$$

onde

$$\begin{aligned} y &= \text{milhas por galão de acordo com o Sindicado dos} \\ &\quad \text{Consumidores com base em testes rodoviários} \\ W &= \text{peso do veículo} \\ S/A &= \text{variável } \textit{dummy} \text{ igual a 0 para transmissão padrão} \\ &\quad \text{e 1 para transmissão automática} \\ G/D &= \text{variável } \textit{dummy} \text{ igual a 0 para veículo a gasolina} \\ &\quad \text{e 1 para veículo a diesel} \\ EPA &= \text{milhagem estimada pela EPA} \end{aligned}$$

Todas as variáveis W, S/A e G/D têm sinais corretos e são significantes, mostrando que a EPA não usou todas as informações disponíveis ao elaborar suas estimativas sobre a eficiência do consumo de combustível [Mad01].

4.2.2 Variáveis *Dummy* para mudanças nos coeficientes angulares

Anteriormente, consideramos variáveis *dummy* para permitir diferenças nos termos de intercepto. Essas variáveis *dummy* assumem valores 0 ou 1. Nem todas as variáveis *dummy* possuem este formato. Também podemos utilizar variáveis *dummy* para permitir diferenças nos coeficientes angulares [Mad01]. Se, por exemplo, as equações de regressão são

$$y_1 = \alpha_1 + \beta_1 x_1 + u_1 \qquad \text{para o primeiro grupo}$$

e

$$y_2 = \alpha_2 + \beta_2 x_2 + u_2 \qquad \text{para o segundo grupo}$$

podemos escrever essas equações juntas como

$$\begin{aligned} y_1 &= \alpha_1 + (\alpha_2 - \alpha_1) \cdot 0 + \beta_1 x_1 + (\beta_2 - \beta_1) \cdot 0 + u_1 \\ y_1 &= \alpha_1 + (\alpha_2 - \alpha_1) \cdot 1 + \beta_1 x_2 + (\beta_2 - \beta_1) \cdot x_2 + u_1 \end{aligned}$$

ou

$$y = \alpha_1 + (\alpha_2 - \alpha_1) D_1 + \beta_1 x + (\beta_2 - \beta_1) D_2 + u \tag{4.22}$$

onde

$$D_1 = \begin{cases} 1 & \text{para todas observações no primeiro grupo} \\ 0 & \text{para todas observações no segundo grupo} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{para todas observações no primeiro grupo} \\ x_2 & \text{isto é, os respectivos valores de } x \text{ para o segundo grupo} \end{cases}$$

O coeficiente de D_1 mede a diferença nos termos de intercepto e o coeficiente de

D_2 mede as diferenças nas inclinações. A estimativa da Equação 4.22 equivale à estimativa de duas equações separadamente se assumirmos que os erros têm uma distribuição idêntica. Excluir D_2 da Equação 4.22 equivale permitir diferentes interceptos mas não diferentes inclinações, e excluir D_1 equivale permitir diferentes inclinações, mas não diferentes interceptos [Mad01].

Variáveis *dummy* apropriadas podem ser definidas quando há mudanças nas inclinações e interceptos em diferentes períodos [Mad01]. Suponha que tenhamos dados para três períodos e, no segundo, apenas o intercepto muda (há deslocamento paralelo). No terceiro período, o intercepto e a inclinação mudam. Então escrevemos

$$\begin{aligned} y_1 &= \alpha_1 + \beta_1 x_1 + u_1 \text{ para o período 1} \\ y_2 &= \alpha_2 + \beta_1 x_2 + u_2 \text{ para o período 2} \\ y_3 &= \alpha_3 + \beta_2 x_3 + u_3 \text{ para o período 3} \end{aligned} \tag{4.23}$$

Logo, podemos combinar essas equações e escrever o modelo como

$$\begin{aligned} y = \alpha_1 &+ (\alpha_2 - \alpha_1) D_1 + (\alpha_3 - \alpha_1) D_2 \\ &+ \beta_1 x + (\beta_2 - \beta_1) D_3 + u \end{aligned} \tag{4.24}$$

onde

$$D_1 = \begin{cases} 1 & \text{para todas observações no período 2} \\ 0 & \text{para outros períodos} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{para todas observações no período 3} \\ 0 & \text{para outros períodos} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{para todas observações nos períodos 1 e 2} \\ x_3 & \text{os valores respectivos de } x \text{ para todas as observações no período 3} \end{cases}$$

Note que em todos estes exemplos estamos assumindo que todos os termos de erro nos diferentes grupos têm a mesma distribuição. Essa é a razão pela qual combinamos os dados de diferentes grupos e escrevemos um tempo de erro u como em 4.22 ou 4.24 e estimamos a equação pelos mínimos quadrados [Mad01].

Uma forma alternativa de escrever as Equações 4.23, é dispor as variáveis y e os termos de erro em colunas [Mad01]. Então escrevemos todos os parâmetros α_1, α_2, α_3, β_1, β_2, com seus fatores multiplicativos dispostos em colunas, da seguinte forma:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \alpha_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_1 \\ x_2 \\ 0 \end{pmatrix} + \beta_2 \begin{pmatrix} 0 \\ 0 \\ x_3 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \tag{4.25}$$

Isso significa que

$$\begin{aligned} y_1 &= \alpha_1(1) + \alpha_2(0) + \alpha_3(0)\beta_1(x_1) + \beta_2(0) + u_1 \\ y_2 &= \alpha_1(0) + \alpha_2(1) + \alpha_3(0)\beta_1(x_2) + \beta_2(0) + u_1 \\ y_3 &= \alpha_1(0) + \alpha_2(0) + \alpha_3(1)\beta_1(0) + \beta_2(x_3) + u_1 \end{aligned}$$

Agora podemos escrever essas equações como

$$y = \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \beta_1 D_4 + \beta_2 D_5 + u \tag{4.26}$$

onde as definições de D_1, D_2, D_3, D_4, D_5 são claramente oriundas da Equação 4.25. Por exemplo,

$$D_2 = \begin{cases} 1 & \text{para todas observações no período 2} \\ 0 & \text{para outros períodos} \end{cases}$$

$$D_4 = \begin{cases} x_i & \text{valores correspondentes de } x \text{ para as observações nos períodos 1 e 2} \\ 0 & \text{para todas observações no período 3} \end{cases}$$

Note que a Equação 4.26 tem que ser estimada sem um termo constante [Mad01].

4.2.3 Variáveis *Dummy* para restrições de equações cruzadas

O método descrito anteriormente pode ser estendido ao caso em que alguns parâmetros dentre as equações são iguais [Mad01]. Como ilustração, considere a estimativa conjunta da demanda por carne bovina, suína e de galinha como base nas séries apresentadas na Tabela 4.10. Waugh [Wau64] estima um conjunto de equações de demanda da forma

$$\begin{aligned} p_1 &= \alpha_1 + \beta_{11}x_1 + \beta_{12}x_2 + \beta_{13}x_3 + \gamma_1 y + u_1 \\ p_1 &= \alpha_2 + \beta_{12}x_1 + \beta_{22}x_2 + \beta_{23}x_3 + \gamma_2 y + u_2 \\ p_1 &= \alpha_3 + \beta_{13}x_1 + \beta_{23}x_2 + \beta_{33}x_3 + \gamma_3 y + u_3 \end{aligned} \tag{4.27}$$

	Carne Bovina		Carne Suína		Carne de Galinha	
Ano	Consumo	Preço por	Consumo	Preço por	Consumo	Preço por
Ano	per Capita	Pound	per Capita	Pound	per Capita	Pound
1948	63,1	82,9	67,8	67,6	18,3	75,4
1949	36,9	76,3	67,7	61,5	19,6	71,8
1950	63,4	88,3	69,2	60,4	20,6	68,0
1951	56,1	90,0	71,9	60,6	21,7	66,0
1952	62,2	85,4	72,4	57,3	22,1	65,0
1953	77,6	66,2	63,5	62,9	21,9	65,8
1954	80,1	64,1	60,0	63,7	22,8	56,4
1955	82,0	63,2	66,8	54,6	21,3	58,7
1956	85,4	60,9	67,3	51,4	24,4	50,4
1957	84,6	63,1	61,1	57,6	25,5	47,6
1958	80,5	72,0	60,2	60,5	28,2	45,8
1959	81,4	73,3	67,6	52,8	28,9	41,4
1960	85,2	70,4	65,2	54,6	28,2	41,4
1961	88,0	68,3	62,2	53,3	30,3	37,0
1962	89,1	69,8	64,0	52,9	30,2	38,6

Tabela 4.10: Consumo *Per Capita* e preços deflacionados [Wau64]

onde

$p_1 =$ preço de varejo da carne bovina

$p_2 =$ preço de varejo da carne suína

$p_3 =$ preço de varejo da carne de galinha

$x_1 =$ consumo *Per Capita* da carne bovina

$x_2 =$ consumo *Per Capita* da carne suína

$x_3 =$ consumo *Per Capita* da carne de galinha

$y =$ renda disponível *Per Capita*

Os preços da Tabela 4.10 são, porém, preços de varejo divididos por um índice de preço ao consumidor. Portanto, multiplicamo-los pelo índice de preço ao consumidor p para obter p_1, p_2 e p_3. Este índice p e a renda disponível y são os seguintes:

Ano	p	y
1948	0,838	1291
1949	0,830	1271
1950	0,838	1369
1951	0,906	1473
1952	0,925	1520
1953	0,932	1582
1954	0,936	1582
1955	0,934	1660
1956	0,947	1742
1957	0,981	1804
1958	1,007	1826
1959	1,015	1904
1960	1,031	1934
1961	1,041	1980
1962	1,054	2052

O sistema de Equações 4.27 possui em especial a simetria nos coeficientes β [Mad01]. Temos

$$\frac{dp_1}{dx_2} = \frac{dp_2}{dx_1} = \beta_{12} \qquad \frac{dp_1}{dx_3} = \frac{dp_3}{dx_1} = \beta_{13} \qquad \frac{dp_2}{dx_3} = \frac{dp_3}{dx_2} = \beta_{23}$$

Logo, há restrições de equações cruzadas nos coeficientes [Mad01]. Se assumirmos que $V(u_1) = V(u_2) = V(u_3)$, podemos minimizar $(\sum u_1^2 + \sum u_2^2 + \sum u_3^2)$, obter as equações normais e estimar os coeficientes de regressão. Este é o método utilizado por Waugh [Wau64]. Ele envolve o desenvolvimento das expressões algébricas necessárias e a programação de tudo novamente. Ao invés disto, podemos utilizar o método das variáveis *dummy*. Podemos escrever as Equações 4.27 como uma única equação

$$\begin{aligned}
\begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} &= \alpha_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + \beta_{11} \begin{pmatrix} x_1 \\ 0 \\ 0 \end{pmatrix} \\
&+ \beta_{12} \begin{pmatrix} x_2 \\ x_1 \\ 0 \end{pmatrix} + \beta_{13} \begin{pmatrix} x_3 \\ 0 \\ x_1 \end{pmatrix} + \beta_{22} \begin{pmatrix} 0 \\ x_2 \\ 0 \end{pmatrix} + \beta_{23} \begin{pmatrix} 0 \\ x_3 \\ x_2 \end{pmatrix} \\
&+ \beta_{33} \begin{pmatrix} 0 \\ 0 \\ x_3 \end{pmatrix} + \gamma_1 \begin{pmatrix} y \\ 0 \\ 0 \end{pmatrix} + \gamma_2 \begin{pmatrix} 0 \\ y \\ 0 \end{pmatrix} + \gamma_3 \begin{pmatrix} 0 \\ 0 \\ y \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}
\end{aligned}$$

Processando as equações, obtemos os seguintes resultados

$$\begin{aligned}
\hat{p}_1 &= 118,98 - 1,534x_1 - 0,474x_2 - 0,445x_3 + 0,0650y \\
\hat{p}_2 &= 149,79 - 0,474x_1 - 1,189x_2 - 0,319x_3 + 0,0162y \\
\hat{p}_3 &= 131,06 - 0,445x_1 - 0,319x_2 - 2,389x_3 + 0,0199y
\end{aligned}$$

5 Descrição do sistema proposto

Neste capítulo será detalhado o estudo das técnicas que serão utilizadas na elaboração do sistema proposto e também a modelagem conceitual do sistema.

5.1 Processo de desenvolvimento

O sistema proposto tem como um de seus objetivos ser adaptável aos diversos processos de desenvolvimento utilizados nos dias de hoje, logo, o método de criação de estimativas utilizado deve ser flexível ao ponto de exigir o mínimo de mudanças no processo de desenvolvimento.

Cada empresa possui seu método de criação de estimativas, adaptado a sua realidade e podendo ser baseados em algum dos diversos métodos existentes na indústria e podendo levar em consideração diversas técnicas de medida de tamanho de software, processos de desenvolvimento, métricas coletadas, etc.

É tarefa do método de criação de estimativas se adaptar a realidade da empresa, ajustando-se ao seu processo de desenvolvimento, e assim, utilizando toda maturidade do processo de desenvolvimento e conhecimento de projetos passados que a empresa possui.

Como vimos anteriormente, não podemos afirmar que um método de estimativas é melhor do que outro em todos os casos. Cada método possui seus pontos positivos e negativos, e baseados nesta afirmação, podemos definir o método a ser utilizado pelo sistema proposto, tendo em vista a necessidade de uma grande flexibilidade.

5.2 Método utilizado

O uso de técnicas baseadas em regressão consistituem a maneira mais comum de criar modelos de custo. A razão para sua popularidade pode ser resumida na sua facilidade de uso e simplicidade [BAC00].

O uso de técnicas baseadas em regressão é apropriado quando:

- Muitos dados estão disponíveis. Isto indica que há vários "graus de liberdade" disponíveis e o número de pontos a serem observados é muito maior do que o numero de variáveis a serem previstas. Coletar dados tem sido um dos maiores desafios neste campo devido a coexistência de vários processos de desenvolvimento e a falta de uma interpretação própria dos processos.

- Nenhum dado está faltando. Dados com informações faltando podem ser notados quando o tempo de coleta e fundos são escassos ou por falta de entendimento dos dados que estão sendo coletados.

- Não existem dados muito discrepantes. Casos extremos são geralmente relatados em dados de engenharia de software, muitas vezes devido a mal entendidos e falta de precisão no processo de coleta de dados.

- As variáveis não estão correlacionadas.

Cada um dos pontos acima mencionados são desafios na modelagem de dados de engenharia de software para desenvolver um modelo de estimativa de custo robusto, simples e construtivo.

Uma das técnicas mais conhecidas de regressão é baseada no método dos quadrados médios mínimos, que foi analisada anteriormente. O método dos quadrados mínimos alivia o problema comum de dados muito discrepantes observados em engenharia de software [BAC00], pois torna fácil a sua identificação e conseqüentemente sua correção. Além deste, existem outros desafios para sua implementação, mas os mesmos são facilmente superados com o uso de um processo de desenvolvimento conciso e confiável.

O método dos quadrados mínimos será utilizado como técnica de regressão, pois o mesmo atende todas as necessidades do sistema proposto, podendo ser utilizando em uma vasta gama de ambientes e adaptar-se a homogeneidade entre os mais diversos processos de desenvolvimento. Esta flexibilidade, e a facilidade no seu uso quando implantado corretamente, são pontos essenciais do sistema proposto.

Mesmo com tais benefícios, devemos ter alguns cuidados na utilização da regressão para gerar estimativas baseadas em dados históricos, pois um pequeno dado discrepante ou uma variável omitida pode comprometer seriamente a qualidade das equações de regressão geradas pelo método. Por este motivo, algumas informações são necessárias, tais como:

- Erro padrão

- Intervalos de Confiança
- Resíduos
- Sinais dos coeficientes de regressão
- Variância

Tais informações devem ser analisadas sempre que um novo conjunto de equações de regressão forem geradas, para garantir sempre que as equações de regressão foram geradas corretamente e conseqüentemente as estimativas geradas pelas mesmas serão confiáveis o bastante para serem utilizadas no projeto.

5.2.1 Dados quantitativos e qualitativos

Durante o decorrer deste trabalho, vimos diversos exemplos de dados quantitativos e qualitativos de projeto que podem ser coletados para o posterior uso na calibragem do modelo gerado. Abaixo citamos alguns exemplos comuns no cenário de desenvolvimento de software atual:

- Tempo real utilizado na execução de uma tarefa
- Tempo estimado para execução de uma tarefa
- Número de Pontos de Função
- Número de cenários de teste
- Número de Casos de Uso
- Linguagem de programação utilizada
- Tipo de software
- Pessoa que realizou a tarefa
- Aplicação legada que recebeu alterações

Estes são só alguns exemplos que podem ser coletados durante a realização do projeto, podendo variar drasticamente entre organizações e processos diferentes, pois dados úteis a um processo, podem não ter utilidade em outro e vice-versa.

Como podemos notar, alguns destes dados não são quantificáveis, são dados puramente qualitativos. Como vimos, em muitos casos, estes dados são de grande importância para uma estimativa, e não podem ser ignorados. Logo, é necessário que levemos em consideração um método de análise de dados qualitativos.

Como vimos anteriormente, variáveis *dummy* podem ser utilizadas para mapear um conjuntos de dados qualitativos em uma equação de regressão. Devido a esta possibilidade e seu uso comprovado, iremos utilizar este método para realizar a análise de dados qualitativos, mas devemos ter cuidado com algumas peculiaridades.

Variáveis *dummy* devem ser tratadas diferentemente pelo sistema, pois devem ser mapeadas diferentemente na equação de regressão. Devido a diferença no modo como se tratam este tipo de variáveis, devemos ter em mente que a nível operacional, o sistema necessita diferenciar este tipo de variáveis das variáveis quantitativas, para que o mesmo possa mapear corretamente estas variáveis na equação de regressão.

Com isto, o sistema proposto pode ser utilizado para estimaras mais variadas tarefas de um projeto, desde a quantidade de pontos de função até quem irá realizar a tarefa. Isto nos dá uma grande flexibilidade no seu uso, necessária ao método implementado pelo sistema proposto.

5.2.2 Modelos diferentes para diferentes tarefas e fases do projeto

Grande vantagem do uso de regressão está na sua flexibilidade, podendo se adaptar a diversos cenários de desenvolvimento e podendo ser utilizado em praticamente todas as fases e para as mais diversas tarefas de um projeto. Tendo em vista este pensamento, devemos salientar alguns aspectos.

O desenvolvimento de software é um processo evolutivo, onde obtemos mais clareza sobre os requisitos conforme o projeto é executado. Lembre do cone da incerteza, onde em fases iniciais do projeto existe um grande grau de incerteza. Este grau de incerteza se reflete em menos informação sobre o projeto. Esta falta de informação sobre o projeto reflete na falta de variáveis explicativas para o modelo de estimativa. Com menos variáveis explicativas, especialmente variáveis muito relevantes, maior será o intervalo de confiança do modelo, o que reflete exatamente o maior grau de incerteza no inicio do projeto.

Conforme o grau de incerteza vai diminuindo, mais variáveis poderão ser utiliza-

das pelo modelo, que conseqüentemente deverá ser capaz de retornar resultados com um intervalo de confiança menor.

Sendo o desenvolvimento de software um processo evolutivo, é natural que possamos realizar diversas estimativas durante sua execução. Como em cada fase do projeto possuí diferentes variáveis explicativas, podemos gerar diversos modelos de estimativa, um para cada fase do projeto onde é necessária a criação de novas estimativas.

Um cuidado deve ser tido conforme atingimos graus de incerteza menores, pois mais variáveis explicativas são geradas, mas nem todas variáveis podem ser relevantes para o nosso modelo. Variáveis explicativas irrelevantes, como visto anteriormente, podem causar problemas em nossas equações de regressão, logo, devemos sempre analisar todas variáveis incluídas para determinar se são realmente significantes.

Outro aspecto a ser notado é de que diferentes tarefas são alteradas por diferentes variáveis. Por exemplo, o número de cenários de teste criados pode não influenciar o tempo de codificação de uma funcionalidade. Tendo em vista esta aspecto, não gostaríamos de adicionar a variável irrelevante "número de cenários de teste" a nossa equação de regressão para estimar o tempo de codificação de uma funcionalidade, pois ela pode afetar nossos resultados de forma negativa, onde provavelmente perderíamos precisão. Mesmo sendo a variável "número de cenários de teste" irrelevante para a estimativa de tempo de codificação de uma funcionalidade, ela pode ser muito significante para a estimativa de tempo de execução do teste desta funcionalidade, logo, não podemos excluí-la completamente.

Tendo em vista este aspecto, podemos deduzir que uma boa alternativa para este problema seria criar diferentes modelos de regressão para diferentes tarefas, onde variáveis significantes de um modelo não afetariam outro modelo onde poderiam ser insignificantes.

A granularidade destes modelos é diretamente ligada com a granularidade que desejamos das estimativas, pois necessitamos um modelo para cada tarefa que desejamos estimar. Caso a estimativa gerada seja uma estimativa inicial, onde desejamos somente saber qual o provável esforço total do projeto, podemos gerar somente um modelo para esta fase inicial.

Esta separação de modelos por tarefas e fases pode ser facilmente determinada utilizando uma WBS de atividade, onde definimos a hierarquia do projeto, baseado em quais fases o mesmo irá possuir e quais tarefas terão de ser executadas em cada uma delas.

Levando em consideração estes aspectos, pode-se assumir que o número de modelos de estimativa que nosso sistema deve possuir depende diretamente do número de

fases que o projeto possui e o número de tarefas que cada fase possui.

5.3 Processo de criação de estimativas

O uso de um processo de criação de estimativas completo é fundamental para o sucesso do sistema, sendo necessário um ciclo recorrente para o contínuo aprimoramento das estimativas. Devido a esta necessidade, o sistema proposto deve implementar este processo repetitivo para garantir o correto uso do método definido.

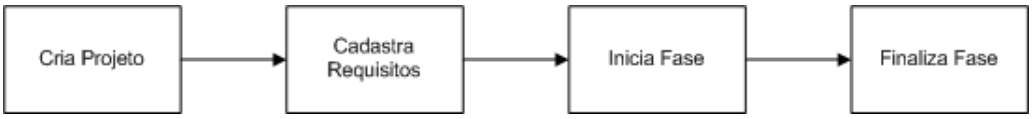


Figura 5.1: Fluxo do processo de criação de estimativas

Neste fluxo pode-se verificar diversas tarefas que deverão ser realizadas durante o processo de criação de uma estimativa. Estas tarefas devem ser implementadas pelo sistema proposto e serão descritas abaixo:

Cria Projeto: Esta tarefa constitui o início do ciclo de estimativas. É o momento onde os detalhes referentes ao projeto como um todo são informados, como por exemplo nome, fases, tarefas, etc...

Cadastra Funcionalidades: Esta tarefa caracteriza o ato de adicionar uma nova funcionalidade ao projeto criado, para que a mesma possa ser estimada conforme as fases e tarefas do projeto.Estas funcionalidades constituem tudo que deve ser estimado no decorrer do projeto. Cada uma destas funcionalidades será estimada separadamente, a partir de seus atributos, para cada uma das tarefas selecionadas em cada fase. Estes atributos devem ser armazenados pelo sistema para o posterior cálculo das estimativas e para geração de equações de regressão mais precisas.

Neste ponto, também deve-se informar que tarefas são pertinentes a funcionalidade em questão. Estas tarefas serão estimadas posteriormente quando a fase ao qual pertencem for iniciada.

Inicia Fase: É o momento onde devem ser informados os atributos de cada funcionalidade que estará sendo estimada na fase em questão. Cada fase necessita de um conjunto específico de dados sobre cada funcionalidade, os atributos necessários para estimativa de cada tarefa da fase.

Assim que informamos os atributos de todas as funcionalidades que estarão sendo estimadas na fase, o sistema deve calcular coeficientes para o modelo de cada tarefa, gerando assim a estimativa para cada funcionalidade do projeto e para todas as tarefas da fase.

Finaliza Fase: Assim que todos as tarefas da fase são executadas, o esforço real utilizado por cada tarefa deve ser informado, para que possa ser utilizado em conjunto com os atributos de cada tarefa para a geração de equações de regressão mais precisas.

Este é um exemplo genérico que pode ser adaptado a diversos processos de desenvolvimento. Mesmo que este processo de estimativa não possa ser utilizado por certo processo de desenvolvimento, isto não significa que a regressão não possa ser utilizada.

É interessante que o sistema proposto também possa nos informar alguns dados sobre os cálculos de regressão, para que possamos identificar eventuais anomalias, como mencionado anteriormente.

5.4 Pacote Flanagan

Na implementação do sistema proposto, será utilizado o pacote Java Flanagan, que implementa uma biblioteca de funções científicas e numéricas escritas pelo Dr. Michael Thomas Flanagan, professor da "University College London" (UCL), para o auxílio em suas pesquisas e projetos [Web08].

Dentre as funções implementadas pelo pacote, temos diversas funções úteis a regressão linear multivariada, e que serão utilizadas na implementação do sistema proposto, em especial a função de regressão linear com intercepto, função que analisamos anteriormente.

A função adapta dados fornecidos a uma função do tipo $y_i = a_0 + a_1 x_0 + a_2 x_1 + \cdots$ utilizando a técnica de regressão e disponibiliza seus resultados. A classe de regressão oferece métodos para obtenção dos seguintes valores:

- Melhor Estimativa dos Coeficientes
- Estimativas dos Desvios Padrão
- Coeficientes de Correlação
- Valores Calculados

- Resíduos
- Soma dos Quadrados dos Resíduos

Mais informações sobre o pacote e o trabalho do Dr. Flanagan podem ser encontrados em "http://www.ee.ucl.ac.uk/∼mflanaga/".

5.5 Modelagem conceitual

Nesta sessão será abordada a modelagem conceitual do sistema proposto. Inicialmente será apresentado um diagrama de casos de uso, sua descrição e logo após, um diagrama de classes.

Tendo em vista que o sistema proposto deve implementar um modelo para cada tarefa do projeto, e cada requisito deve possuir os atributos necessários a estes modelos, podemos concluir que o sistema proposto depende diretamente do processo de desenvolvimento utilizado por ele. Sendo assim, o mesmo deve implementar os modelos e atributos necessários ao processo. Devido a este fato, a modelagem conceitual do sistema deve ser feita com base em um processo de desenvolvimento pré-determinado.

Para que os testes possam ser realizados, um processo simples foi definido, contendo somente duas fases, "Development" e "Testing", cada uma contendo uma tarefa. Este processo foi definido tendo em vista uma realidade onde cada projeto visa a manutenção de aplicações legadas e não a criação de novas aplicações.

A primeira fase possui a tarefa de "Coding and Unit Test", que necessita por sua vez, dois atributos, a aplicação que está sendo alterada pela funcionalidade e sua estimativa de esforço utilizando a técnica Delphi. A segunda fase, possui a tarefa de "SIT Execution", e possui três atributos, novamente a aplicação que está sendo alterada, sua estimativa de esforço utilizando a técnica Delphi e o número de cenários de testes a serem executados. Podemos notar que temos dados quantitativos e qualitativos em ambas tarefas, que devem ser tratados pelos respectivos modelos para a criação das estimativas.

5.5.1 Diagrama de casos de uso

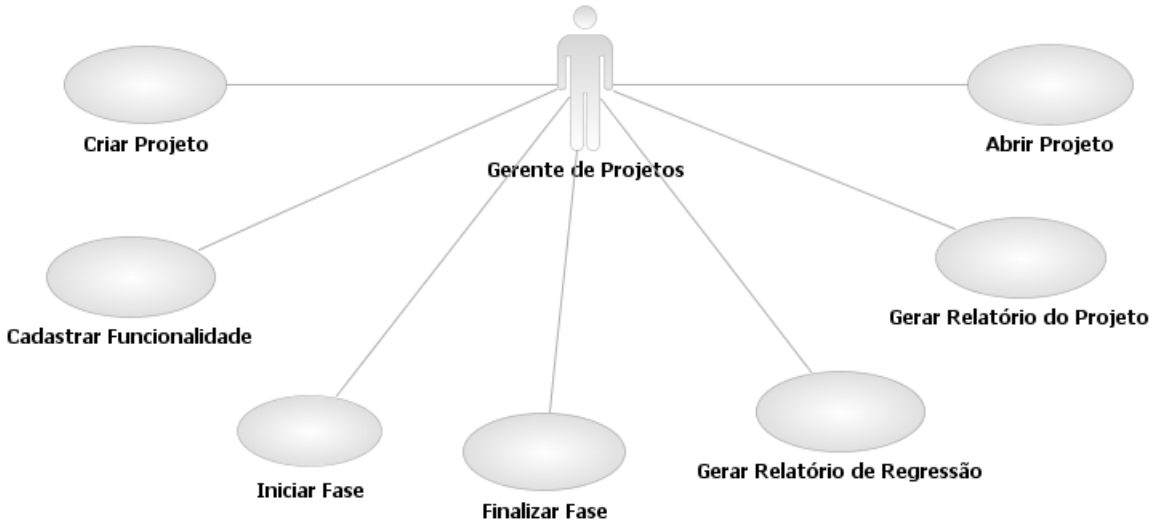


Figura 5.2: Diagrama de casos de uso

5.5.1.1 Descrição de casos de uso

Identificador:	UC001	**Nome:**	Criar Projeto
Ator:	Gerente de projeto		
Descrição:	Esta ação constitui a criação de um novo projeto no sistema.		
Pré-Condições:	NA		
Pós-Condições:	Projeto criado		
Curso normal de eventos:	1. Gerente de projetos requisita criação de um novo projeto. 2. Gerente de projetos informa o nome do projeto. 3. Sistema cria novo projeto.		
Curso alternativo de eventos:	NA		
Exceções:	NA		

Identificador:	UC002	**Nome:**	Cadastra Funcionalidade
Ator:	Gerente de projeto		
Descrição:	Esta ação constitui o cadastro de uma nova funcionalidade em um projeto já criado.		
Pré-Condições:	Projeto aberto		
Pós-Condições:	Funcionalidade cadastrada		
Curso normal de eventos:	1. Gerente de projetos requisita o cadastro de uma nova funcionalidade. 2. Gerente de projetos informa o nome da funcionalidade. 3. Gerente de projetos seleciona as tarefas necessárias a funcionalidade. 4. Sistema cadastra nova funcionalidade no projeto. 5. Sistema adiciona as tarefas selecionadas para cada funcionalidade.		
Curso alternativo de eventos:	**Nenhum projeto aberto** 2. Sistema informa que não há projeto aberto.		
Exceções:	NA		

Identificador:	UC003	**Nome:**	Inicia Fase
Ator:	Gerente de projeto		
Descrição:	Esta ação constitui o inicio de uma nova fase do projeto.		
Pré-Condições:	Projeto aberto Ao menos uma funcionalidade cadastrada Ao menos uma fase para ser iniciada		
Pós-Condições:	Fase iniciada Dados informados pelo gerente de projetos armazenados Estimativas para as tarefas da fase geradas		
Curso normal de eventos:	1. Gerente de projetos requisita o inicio de uma fase. 2. Gerente de projetos seleciona qual fase deseja iniciar. 3. Gerente de projetos informa dados sobre cada tarefa estimada para a fase. 4. Sistema armazena dados informados pelo gerente de projetos. 5. Sistema gera estimativas para as tarefas baseado nas informações de cada funcionalidade e equações de regressão. 6. Sistema marca fase como iniciada.		
Curso alternativo de eventos:	**Nenhum projeto aberto** 2. Sistema informa que não há projeto aberto. **Nenhuma fase para ser iniciada** 2. Sistema informa que não há fases para serem iniciadas.		
Exceções:	NA		

Identificador:	UC004	**Nome:**	Finaliza Fase
Ator:	Gerente de projeto		
Descrição:	Esta ação constitui o fim de uma fase do projeto.		
Pré-Condições:	Projeto aberto Fase iniciada		
Pós-Condições:	Fase finalizada Esforço real de cada tarefa armazenado.		
Curso normal de eventos:	1. Gerente de projetos requisita o fim de uma fase. 2. Gerente de projetos informa o esforço real de cada tarefa realizada. 3. Sistema armazena o esforço real de cada tarefa realizada. 4. Sistema marca fase como finalizada.		
Curso alternativo de eventos:	**Nenhum projeto aberto** 2. Sistema informa que não há projeto aberto. **Nenhuma fase para ser finalizada** 2. Sistema informa que não há fases para serem finalizadas.		
Exceções:	NA		

Identificador:	UC005	**Nome:**	Gera relatório de regressão
Ator:	Gerente de projeto		
Descrição:	Esta ação constitui a geração de um relatório com dados sobre a regressão de um modelo, para que se possa analisar possíveis problemas com a regressão		
Pré-Condições:	Existe ao menos um modelo criado		
Pós-Condições:	Relatório de regressão gerado		
Curso normal de eventos:	1. Gerente de projetos requisita a geração do relatório. 2. Gerente de projetos seleciona a fase que possui a tarefa desejada. 3. Gerente de projetos seleciona a tarefa a ser analisada. 4. Sistema exibe o relatório de regressão do modelo contendo os coeficientes de regressão, coeficiente de correlação linear, funcionalidades analisadas, resíduos, erros, e média de erro.		
Curso alternativo de eventos:	NA		
Exceções:	NA		

Identificador:	UC006	**Nome:**	Gera relatório do projeto
Ator:	Gerente de projeto		
Descrição:	Esta ação constitui a geração de um relatório com dados sobre um projeto, para que se possa analisar o andamento do projeto.		
Pré-Condições:	Projeto aberto		
Pós-Condições:	Relatório de projeto gerado		
Curso normal de eventos:	1. Gerente de projetos requisita a geração do relatório. 2. Sistema exibe o relatório de projeto contendo o nome do projeto, fases iniciadas, fases finalizadas, tarefas cadastradas, funcionalidades adicionadas, esforço estimado, esforço real		
Curso alternativo de eventos:	**Nenhum projeto aberto** 2. Sistema informa que não há projeto aberto.		
Exceções:	NA		

Identificador:	UC007	**Nome:**	Abrir Projeto
Ator:	Gerente de projeto		
Descrição:	Esta ação constitui a seleção de um projeto já criado no sistema.		
Pré-Condições:	Ao menos um projeto criado		
Pós-Condições:	Projeto aberto		
Curso normal de eventos:	1. Gerente de projetos requisita a abertura de um projeto. 2. Gerente de projetos seleciona o nome do projeto. 3. Sistema abre o projeto.		
Curso alternativo de eventos:	**Nenhum projeto criado** 2. Sistema informa que não há projetos criados.		
Exceções:	NA		

5.5.2 Diagrama de classes

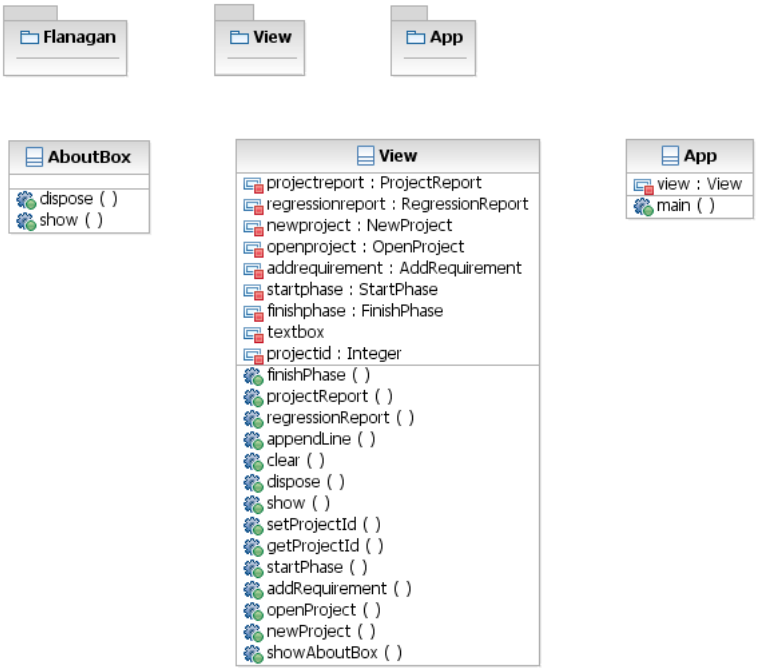


Figura 5.3: Diagrama de classes

5.5.2.1 Pacote Flanagan

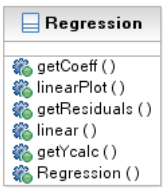


Figura 5.4: Diagrama de classes: Pacote Flanagan

5.5.2.2 Pacote View

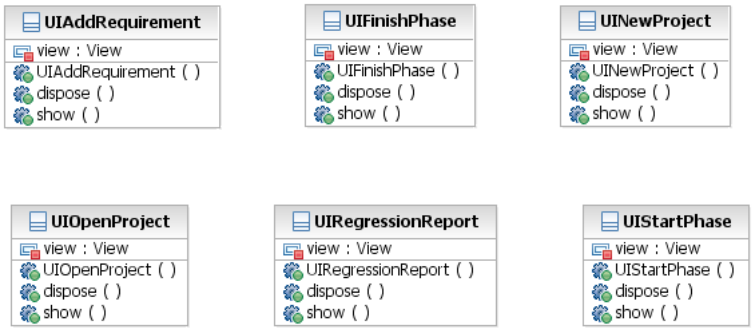


Figura 5.5: Diagrama de classes: Pacote View

5.5.2.3 Pacote App

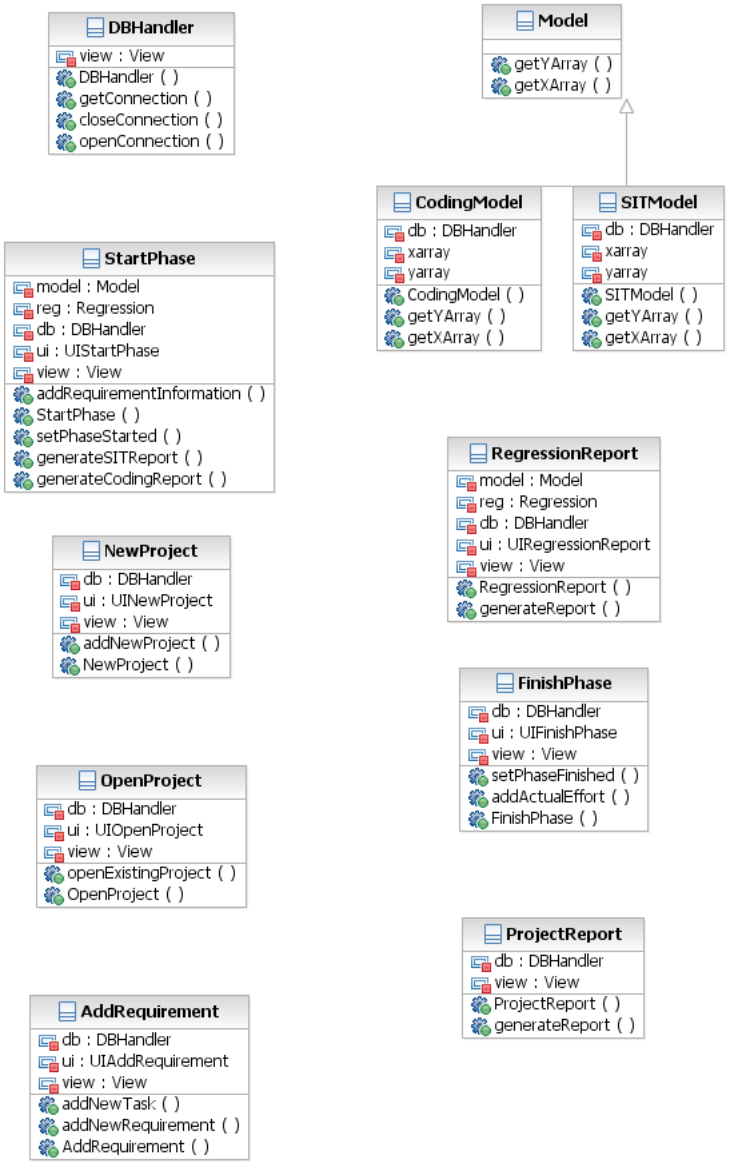


Figura 5.6: Diagrama de classes: Pacote App

5.5.3 Diagrama ER

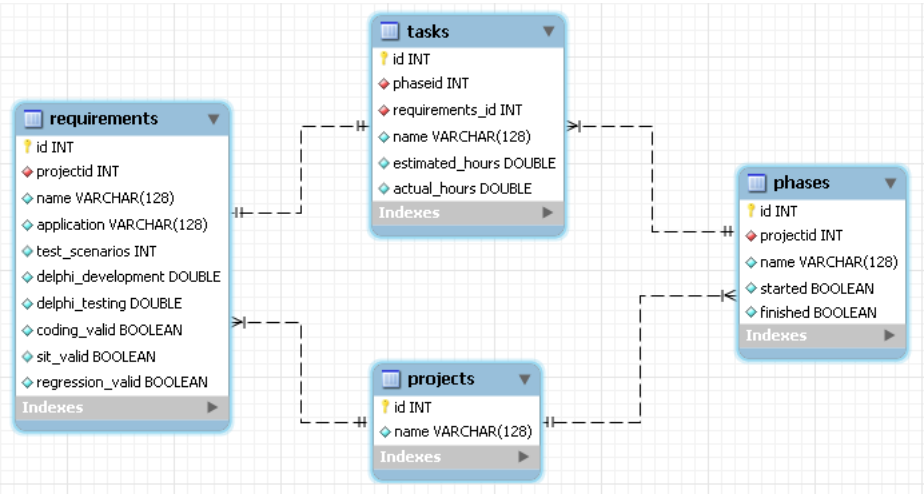


Figura 5.7: Diagrama ER

5.6 Recursos necessários

Nesta sessão serão apresentados os recursos de tecnologia necessários à realização deste trabalho.

5.6.1 Recursos de software

Os recursos de software necessários para realização do trabalho são:

- Editor LaTex para Windows: TeXnicCenter 1 Beta 7.01
- Engine LaTex para Windows: MiKTeX 2.6
- Padrão de formatação ABNT: abnTeX 0.8-2
- Sistema para organização de referências: JabRef 2.1b2
- Ferramenta de modelagem: Rational Software Architect 7.0 da Rational Software Corporation
- Visualizador de documentos: Acrobat Professional da Adobe Systems Incorporated
- Navegador Web: Mozilla Firefox da Mozilla Foundation
- Software de análise de dados estatísticos: SPSS 16.0 for Windows da Statistical Product and Service Solutions (SPSS)
- Classes de desenvolvimento: Java Software Development Kit 1.6.0 Update 6 da Sun Microsystems
- Ambiente de desenvolvimento: NetBeans 6.1
- Biblioteca Java: Dr. Michael Thomas Flanagan's Java Scientific Library - UCL (University College London)
- Banco de Dados: MySQL 5.0 da MySQL AB
- Ferramenta de gerência de Banco de Dados: MySQL GUI Tools 5.0 da MySQL AB
- Ferramenta de design de Banco de Dados: MySQL Workbench 5.0.22 da MySQL AB

5.6.2 Recursos de hardware

Os recursos de hardware deste projeto foram definidos pelas necessidades dos recursos de software.

Estação de trabalho com acesso a internet e a seguinte configuração mínima:

- Sistema operacional Windows XP Professional, Service Pack 2
- Processador recomendado: 1GHz
- Memória RAM disponível: 512Mb
- Espaço livre em disco: 3Gb

6 Testes realizados e resultados obtidos

Neste capítulo serão abordados os testes realizados sobre a solução proposta anteriormente. Para a realização dos testes, o sistema descrito no capítulo anterior foi implementado utilizando a linguagem Java versão 1.6.0 Update 6 e sua documentação pode ser encontrada no Apêndice B.

6.1 Contextualização

O sistema implementado assume uma realidade distinta, com um processo de desenvolvimento definido para um tipo de projeto específico. Os projetos em questão são projetos de manutenção de aplicações legadas, onde um conjunto de novas funcionalidades ou correções de problemas de diversas aplicações são agrupados, gerando assim, um novo projeto. Cada projeto possui um escopo preliminar, um cronograma definido e um custo máximo. Uma análise mais aprofundada é feita sobre o escopo preliminar, onde diversas informações sobre cada funcionalidade são obtidas. A partir desta análise as estimativas do projeto são criadas, e a partir delas, o escopo final do projeto é definido e seu desenvolvimento é iniciado.

Baseado nesta análise, analistas e desenvolvedores devem obter as informações necessárias para a definição dos atributos que serão utilizados pelo sistema.

Como mencionamos anteriormente, o sistema implementado deve auxiliar na estimativa de duas tarefas do projeto, “Coding and Unit Test” e “SIT Execution”. A primeira tarefa possui dois atributos, a aplicação que está sendo alterada pela funcionalidade e sua estimativa de esforço utilizando a técnica Delphi. A segunda tarefa possui três atributos, novamente a aplicação que está sendo alterada, sua estimativa de esforço utilizando a técnica Delphi e o número de cenários de testes a serem executados.

Baseados na análise feita anteriormente sobre os requisitos, podere-se obter os atributos de cada requisito, que serão utilizados pelo sistema para geração das estimativas. A aplicação sendo alterada é obtida na definição do requisito, suas estimativas iniciais de

tempo de desenvolvimento e teste são feitas por analistas e desenvolvedores utilizando a técnica Delphi. O número de cenários de teste é definido por analistas de teste, que deverão apontar quais pontos do sistema serão testados. Todos estes atributos serão informados ao sistema deverão ser informados ao sistema, para que o mesmo possa gerar as estimativas para cada tarefa.

6.2 Testes realizados

Para a realização dos testes, necessitamos de uma carga de dados inicial, para que as equações de regressão possam ser geradas. Um conjunto de dados reais de projetos acima mencionados foi utilizado para popular o sistema. Inicialmente adicionamos ao sistema um projeto, denominado "Projeto 1", possuindo 108 funcionalidades, ou requisitos, cada um contendo os atributos necessários a todas as tarefas do projeto. Estes requisitos correspondem a dados reais de projetos, e foram obtidos a partir da uma análise do escopo, como mencionado anteriormente. Os requisitos populados inicialmente podem ser vistos na Tabela 6.1.

Requisito	**Aplicação**	**Desenvolvimento (Delphi)**	**Teste (Delphi)**	**Cenários de Teste**
Req 1	App 1	25,03	26,8	43
Req 2	App 1	7,38	14	9
Req 3	App 1	53,72	7	14
Req 4	App 1	10	5,8	3
Req 5	App 1	16,83	17,5	6
Req 6	App 2	3,17	3,5	41
Req 7	App 2	3,67	2,33	
Req 8	App 2	21,08	9,92	76
Req 9	App 2	4,17	1,17	4
Req 10	App 2	18,67	2,33	9
Req 11	App 2	5,67	2,92	10
Req 12	App 2	4,58	2,33	10
Req 13	App 2	18,67	4,67	16
Req 14	App 2	18,67	3,5	21
Req 15	App 2	7,58	4,67	7
Req 16	App 2	3,08	0,58	5
Req 17	App 2	16,5	0,58	4
Req 18	App 2	4,58	3,5	3
Req 19	App 2	19,67	1,75	3
Req 20	App 2	9,17	1,75	2
Req 21	App 2	22,92	2,33	7
Req 22	App 2	16,5	4,67	6
Req 23	App 2	32	4,67	4
Req 24	App 2	33,83	4,08	4
Req 25	App 2	33,83	1,75	4
Req 26	App 2	3,08	1,75	4
Req 27	App 2	6,58	1,75	7

Continua na próxima página

Requisito	Aplicação	Desenvolvimento (Delphi)	Teste (Delphi)	Cenários de Teste
Req 28	App 2	3,17	1,17	8
Req 29	App 2	3,17	1,75	4
Req 30	App 2	3,08	0,58	3
Req 31	App 2	9,08	2,33	9
Req 32	App 2	9,08	2,33	7
Req 33	App 2	22,17	5,83	7
Req 34	App 2	3,17	3,5	2
Req 35	App 2	11,08	3,5	5
Req 36	App 2	14	4,67	4
Req 37	App 2	14	4,67	4
Req 38	App 2	14	4,67	4
Req 39	App 2	23,5	7	5
Req 40	App 2	14	5,83	10
Req 41	App 2	14	3,5	7
Req 42	App 2	14	4,67	4
Req 43	App 2	14	4,67	5
Req 44	App 3	30	14	5
Req 45	App 3	62	14	4
Req 46	App 3	40	23,33	16
Req 47	App 3	20	14	3
Req 48	App 4	32	8,17	7
Req 49	App 4	32	5,83	13
Req 50	App 4	8	5,8	5
Req 51	App 4	16	3,5	8
Req 52	App 5	8	5,83	7
Req 53	App 5	40	2,33	37
Req 54	App 5	3,5	4,67	9
Req 55	App 5	8	4,67	27
Req 56	App 5	10	9,33	3
Req 57	App 5	3	2,33	3
Req 58	App 5	6	4,67	7
Req 59	App 5	12	4,67	11
Req 60	App 5	10	9,33	33
Req 61	App 5	8	4,67	10
Req 62	App 5	8	3,5	7
Req 63	App 5	6	3,5	3
Req 64	App 6	6,3	2,33	5
Req 65	App 6	16	7	7
Req 66	App 6	25,4	4,67	6
Req 67	App 6	16	2,33	4
Req 68	App 6	25,4	9,33	6
Req 69	App 6	25,4	7	5
Req 70	App 6	10	4,67	2
Req 71	App 7	16	4,7	1
Req 72	App 8	80	1,2	
Req 73	App 9	28	4,67	2
Req 74	App 9	73	35	13
Req 75	App 9	40	7	4
Req 76	App 9	21	7	5
Req 77	App 9	28	4,67	2
Req 78	App 9	8	9,33	16
Req 79	App 9	4	2,33	9
Req 80	App 10	9,98	6,65	16
Req 81	App 10	1	10	16
Req 82	App 10	52,05	12	12

Continua na próxima página

Requisito	Aplicação	Desenvolvimento (Delphi)	Teste (Delphi)	Cenários de Teste
Req 83	App 10	14	11	30
Req 84	App 1	4	7	21
Req 85	App 1	1	23	110
Req 86	App 1	1,98	9	10
Req 87	App 11	57,48	50	172
Req 88	App 11	5,8	8	18
Req 89	App 11	1,2	3	4
Req 90	App 12	0,98	6	28
Req 91	App 9	14,1	42	37
Req 92	App 9	9,48	39	40
Req 93	App 12	24	4	6
Req 94	App 12	24	3	6
Req 95	App 2	42	20,98	69
Req 96	App 2	24	4,98	17
Req 97	App 2	25	5	7
Req 98	App 2	26	14	11
Req 99	App 11	40	9,98	27
Req 100	App 11	6	5	24
Req 101	App 13	100	12	5
Req 102	App 13	21	15,5	10
Req 103	App 13	84	20	22
Req 104	App 13	10	10	9
Req 105	App 10	110	34,98	88
Req 106	App 10	60	9	14
Req 107	App 10	40	13,98	14
Req 108	App 1	21	25	65

Tabela 6.1: Requisitos populados no “Projeto 1”

Também adicionamos ao sistema as tarefas realizadas no “Projeto 1”, juntamente com o esforço necessário para realizá-las. Estas tarefas podem ser vistas na Tabela 6.2

Requisito	Tarefa	Esforço	Tarefa	Esforço
Req 1	Coding and Unit Test	57,4	SIT Execution	44,3
Req 2	Coding and Unit Test	21,7	SIT Execution	7,5
Req 3	Coding and Unit Test	61,6	SIT Execution	10
Req 4	Coding and Unit Test	15	SIT Execution	7,5
Req 5	Coding and Unit Test	29,2	SIT Execution	7
Req 6	Coding and Unit Test	13,5	SIT Execution	13,6
Req 7	Coding and Unit Test	3,8	SIT Execution	0
Req 8	Coding and Unit Test	48,1	SIT Execution	12,3
Req 9	Coding and Unit Test	2,5	SIT Execution	3
Req 10	Coding and Unit Test	4,8	SIT Execution	2,5
Req 11	Coding and Unit Test	8,8	SIT Execution	5
Req 12	Coding and Unit Test	6	SIT Execution	2,5
Req 13	Coding and Unit Test	5	SIT Execution	2
Req 14	Coding and Unit Test	6,5	SIT Execution	2
Req 15	Coding and Unit Test	3,8	SIT Execution	2
Req 16	Coding and Unit Test	1	SIT Execution	0,8
Req 17	Coding and Unit Test	6	SIT Execution	1
Req 18	Coding and Unit Test	41,5	SIT Execution	5,8
Req 19	Coding and Unit Test	28,6	SIT Execution	0
Req 20	Coding and Unit Test	2	SIT Execution	2
Req 21	Coding and Unit Test	9,8	SIT Execution	2
Req 22	Coding and Unit Test	20,9	SIT Execution	3,5
Req 23	Coding and Unit Test	18,6	SIT Execution	3,5
Req 24	Coding and Unit Test	31,2	SIT Execution	2
Req 25	Coding and Unit Test	7,8	SIT Execution	2,5
Req 26	Coding and Unit Test	1	SIT Execution	4,5

Continua na próxima página

Requisito	Tarefa	Esforço	Tarefa	Esforço
Req 27	Coding and Unit Test	1	SIT Execution	0,5
Req 28	Coding and Unit Test	1	SIT Execution	0,5
Req 29	Coding and Unit Test	1	SIT Execution	0,5
Req 30	Coding and Unit Test	0,2	SIT Execution	1,8
Req 31	Coding and Unit Test	17,2	SIT Execution	2
Req 32	Coding and Unit Test	8,8	SIT Execution	3,5
Req 33	Coding and Unit Test	30,9	SIT Execution	5
Req 34	Coding and Unit Test	3	SIT Execution	3,5
Req 35	Coding and Unit Test	8,8	SIT Execution	4,5
Req 36	Coding and Unit Test	4	SIT Execution	4
Req 37	Coding and Unit Test	5	SIT Execution	4
Req 38	Coding and Unit Test	5	SIT Execution	3,5
Req 39	Coding and Unit Test	30,2	SIT Execution	6
Req 40	Coding and Unit Test	24,4	SIT Execution	5
Req 41	Coding and Unit Test	7,8	SIT Execution	2
Req 42	Coding and Unit Test	3	SIT Execution	3
Req 43	Coding and Unit Test	8	SIT Execution	3
Req 44	Coding and Unit Test	35	SIT Execution	16,5
Req 45	Coding and Unit Test	74,63	SIT Execution	9,3
Req 46	Coding and Unit Test	59,5	SIT Execution	12,4
Req 47	Coding and Unit Test	20	SIT Execution	5,5
Req 48	Coding and Unit Test	47,8	SIT Execution	5
Req 49	Coding and Unit Test	32	SIT Execution	8,8
Req 50	Coding and Unit Test	8,8	SIT Execution	3,5
Req 51	Coding and Unit Test	13,6	SIT Execution	3
Req 52	Coding and Unit Test	7,8	SIT Execution	16,1
Req 53	Coding and Unit Test	52,1	SIT Execution	53,7
Req 54	Coding and Unit Test	3,5	SIT Execution	21,8
Req 55	Coding and Unit Test	6,8	SIT Execution	21,3
Req 56	Coding and Unit Test	10	SIT Execution	15
Req 57	Coding and Unit Test	2	SIT Execution	11,5
Req 58	Coding and Unit Test	4	SIT Execution	9,25
Req 59	Coding and Unit Test	5,8	SIT Execution	14,5
Req 60	Coding and Unit Test	10,4	SIT Execution	19,27
Req 61	Coding and Unit Test	21,3	SIT Execution	29
Req 62	Coding and Unit Test	8,8	SIT Execution	14,5
Req 63	Coding and Unit Test	3	SIT Execution	11,5
Req 64	Coding and Unit Test	8,8	SIT Execution	4
Req 65	Coding and Unit Test	16,8	SIT Execution	12,5
Req 66	Coding and Unit Test	41,2	SIT Execution	8,02
Req 67	Coding and Unit Test	18,1	SIT Execution	9,3
Req 68	Coding and Unit Test	21,6	SIT Execution	6,8
Req 69	Coding and Unit Test	70,4	SIT Execution	4,8
Req 70	Coding and Unit Test	10	SIT Execution	2,3
Req 71	Coding and Unit Test	24,7	SIT Execution	2,8
Req 72	Coding and Unit Test	55	SIT Execution	0
Req 73	Coding and Unit Test	163,4	SIT Execution	3
Req 74	Coding and Unit Test	127	SIT Execution	33,2
Req 75	Coding and Unit Test	33,9	SIT Execution	3,5
Req 76	Coding and Unit Test	45,6	SIT Execution	4,5
Req 77	Coding and Unit Test	20,6	SIT Execution	3
Req 78	Coding and Unit Test	3,8	SIT Execution	7,5
Req 79	Coding and Unit Test	4,4	SIT Execution	2,5
Req 80	Coding and Unit Test	28,98	SIT Execution	9,5
Req 81	Coding and Unit Test	1	SIT Execution	23,8
Req 82	Coding and Unit Test	0	SIT Execution	5,25

Continua na próxima página

Requisito	Tarefa	Esforço	Tarefa	Esforço
Req 83	Coding and Unit Test	1,5	SIT Execution	10,75
Req 84	Coding and Unit Test	0	SIT Execution	27,9
Req 85	Coding and Unit Test	49,2	SIT Execution	46,1
Req 86	Coding and Unit Test	0	SIT Execution	9
Req 87	Coding and Unit Test	55	SIT Execution	58,3
Req 88	Coding and Unit Test	35,2	SIT Execution	8
Req 89	Coding and Unit Test	8,8	SIT Execution	3,3
Req 90	Coding and Unit Test	13,1	SIT Execution	6
Req 91	Coding and Unit Test	8,3	SIT Execution	31,6
Req 92	Coding and Unit Test	50,2	SIT Execution	47,7
Req 93	Coding and Unit Test	22,4	SIT Execution	3
Req 94	Coding and Unit Test	22,4	SIT Execution	4
Req 95	Coding and Unit Test	36,2	SIT Execution	11
Req 96	Coding and Unit Test	13,6	SIT Execution	6
Req 97	Coding and Unit Test	26,9	SIT Execution	6
Req 98	Coding and Unit Test	15,8	SIT Execution	7
Req 99	Coding and Unit Test	40,2	SIT Execution	13,6
Req 100	Coding and Unit Test	8,8	SIT Execution	5,4
Req 101	Coding and Unit Test	123,2	SIT Execution	13,5
Req 102	Coding and Unit Test	45,2	SIT Execution	8,6
Req 103	Coding and Unit Test	66,4	SIT Execution	18,8
Req 104	Coding and Unit Test	6	SIT Execution	5
Req 105	Coding and Unit Test	107,6	SIT Execution	36,75
Req 106	Coding and Unit Test	33	SIT Execution	10,25
Req 107	Coding and Unit Test	63,05	SIT Execution	17,25
Req 108	Coding and Unit Test	21	SIT Execution	25,4

Tabela 6.2: Tarefas populadas no “Projeto 1”

Tendo em vista que dados reais de projetos são muitas vezes incompletos ou incorretos, devemos realizar uma análise inicial sobre os dados populados a fim de encontrar possíveis requisitos com este problema.

Nos dados do “Projeto 1” encontramos alguns requisitos que não possuíam informação sobre o número de cenários de teste, são estes, os requisitos “Req 7” e “Req 72”. Também encontramos algumas tarefas cujo esforço real é zero, são estas “Coding and Unit Test” para “Req 82”, “Req 84” e “Req 86”, e “SIT Execution” para “Req 7”, “Req 19” e “Req 72”.

Para que estes requisitos não influenciem o modelo de regressão, devemos descartá-los no cálculo das equações de regressão. Devido a isto, devem ser descartados os requisitos “Req 7”, “Req 19” e “Req 72” para o modelo “SIT Execution” e os requisitos “Req 82”, “Req 84” e “Req 86” para o modelo “Coding and Unit Test”.

Uma vez descartados os requisitos inválidos, podemos gerar o relatório de regressão, conforme citado no Apêndice B, pela funcionalidade “Regression Report”, a fim de encontrar dados discrepantes.

O relatório de regressão é gerado pelo sistema utilizando primariamente funções do pacote Flanagan, que tem como entrada dois conjuntos de dados, uma matriz de variáveis independentes e um *array* de variáveis dependentes. Ambos conjuntos de dados devem ser numéricos, o que nos impõe um problema, pois em nossos modelos de regressão

possuímos uma variável qualitativa, a aplicação sendo alterada. Esta variável deve ser modificada para que possa ser utilizada pelo sistema. Como vimos anteriormente, utilizaremos variáveis *Dummy* para mapear variáveis qualitativas do sistema, transformando cada aplicação em um novo atributo binário, onde o mesmo receberá valor verdadeiro caso o requisito altere a aplicação e falso caso contrário.

Esta abordagem é muito interessante, mas devemos ter cuidado com um pequeno detalhe, não devemos gerar estimativas além do intervalo dos dados utilizados na regressão. Tendo em vista que cada variável *Dummy* representa um diferente intervalo de dados, contemplando todos requisitos cujo valor é verdadeiro, uma variável que possui poucas ocorrências limita muito nosso intervalo de estimativa. Devido a este problema, o sistema anula todas as variáveis *Dummy* que possuam menos de N ocorrências, fazendo com que algumas aplicações sejam ignoradas nos cálculos do modelo. Inicialmente definimos que todas aplicações, ou variáveis *Dummy* com menos de 15 ocorrências não deverão ser utilizadas.

Tendo em vista estas características, o sistema nos gera o seguinte relatório de regressão para a tarefa "Coding and Unit Test":

```
Unweighted Least Squares Minimisation
Linear Regression with intercept
y = c[0] + c[1]*x1 + c[2]*x2 +c[3]*x3 + . . .

        Best Estimates
c[0]    9,7605
c[1]    -11,0845
c[2]    0,9483

Correlation: x - y data
Linear Correlation Coefficient (R): 0,7549
Linear Correlation Coefficient Squared (R^2): 0,5699

x1      x2      y(expl) y(calc) residual
0,0000  25,0300 57,4000 33,4958 23,9042
0,0000  7,3800  21,7000 16,7588 4,9412
0,0000  53,7200 61,6000 60,7017 0,8983
0,0000  10,0000 15,0000 19,2432 -4,2432
0,0000  16,8300 29,2000 25,7199 3,4801
1,0000  3,1700  13,5000 1,6820  11,8180
1,0000  3,6700  3,8000  2,1562  1,6438
1,0000  21,0800 48,1000 18,6656 29,4344
1,0000  4,1700  2,5000  2,6303  -0,1303
1,0000  18,6700 4,8000  16,3803 -11,5803
1,0000  5,6700  8,8000  4,0527  4,7473
1,0000  4,5800  6,0000  3,0191  2,9809
1,0000  18,6700 5,0000  16,3803 -11,3803
1,0000  18,6700 6,5000  16,3803 -9,8803
1,0000  7,5800  3,8000  5,8639  -2,0639
1,0000  3,0800  1,0000  1,5967  -0,5967
1,0000  16,5000 6,0000  14,3225 -8,3225
1,0000  4,5800  41,5000 3,0191  38,4809
1,0000  19,6700 28,6000 17,3285 11,2715
1,0000  9,1700  2,0000  7,3717  -5,3717
1,0000  22,9200 9,8000  20,4104 -10,6104
1,0000  16,5000 20,9000 14,3225 6,5775
1,0000  32,0000 18,6000 29,0207 -10,4207
```

1,0000	33,8300	31,2000	30,7561	0,4439
1,0000	33,8300	7,8000	30,7561	−22,9561
1,0000	3,0800	1,0000	1,5967	−0,5967
1,0000	6,5800	1,0000	4,9157	−3,9157
1,0000	3,1700	1,0000	1,6820	−0,6820
1,0000	3,1700	1,0000	1,6820	−0,6820
1,0000	3,0800	0,2000	1,5967	−1,3967
1,0000	9,0800	17,2000	7,2863	9,9137
1,0000	9,0800	8,8000	7,2863	1,5137
1,0000	22,1700	30,9000	19,6992	11,2008
1,0000	3,1700	3,0000	1,6820	1,3180
1,0000	11,0800	8,8000	9,1829	−0,3829
1,0000	14,0000	4,0000	11,9518	−7,9518
1,0000	14,0000	5,0000	11,9518	−6,9518
1,0000	14,0000	5,0000	11,9518	−6,9518
1,0000	23,5000	30,2000	20,9604	9,2396
1,0000	14,0000	24,4000	11,9518	12,4482
1,0000	14,0000	7,8000	11,9518	−4,1518
1,0000	14,0000	3,0000	11,9518	−8,9518
1,0000	14,0000	8,0000	11,9518	−3,9518
0,0000	30,0000	35,0000	38,2087	−3,2087
0,0000	62,0000	74,6300	68,5534	6,0766
0,0000	40,0000	59,5000	47,6914	11,8086
0,0000	20,0000	20,0000	28,7260	−8,7260
0,0000	32,0000	47,8000	40,1052	7,6948
0,0000	32,0000	32,0000	40,1052	−8,1052
0,0000	8,0000	8,8000	17,3467	−8,5467
0,0000	16,0000	13,6000	24,9329	−11,3329
0,0000	8,0000	7,8000	17,3467	−9,5467
0,0000	40,0000	52,1000	47,6914	4,4086
0,0000	3,5000	3,5000	13,0795	−9,5795
0,0000	8,0000	6,8000	17,3467	−10,5467
0,0000	10,0000	10,0000	19,2432	−9,2432
0,0000	3,0000	2,0000	12,6053	−10,6053
0,0000	6,0000	4,0000	15,4502	−11,4502
0,0000	12,0000	5,8000	21,1398	−15,3398
0,0000	10,0000	10,4000	19,2432	−8,8432
0,0000	8,0000	21,3000	17,3467	3,9533
0,0000	8,0000	8,8000	17,3467	−8,5467
0,0000	6,0000	3,0000	15,4502	−12,4502
0,0000	6,3000	8,8000	15,7346	−6,9346
0,0000	16,0000	16,8000	24,9329	−8,1329
0,0000	25,4000	41,2000	33,8466	7,3534
0,0000	16,0000	18,1000	24,9329	−6,8329
0,0000	25,4000	21,6000	33,8466	−12,2466
0,0000	25,4000	70,4000	33,8466	36,5534
0,0000	10,0000	10,0000	19,2432	−9,2432
0,0000	16,0000	24,7000	24,9329	−0,2329
0,0000	80,0000	55,0000	85,6223	−30,6223
0,0000	28,0000	163,4000	36,3121	127,0879
0,0000	73,0000	127,0000	78,9843	48,0157
0,0000	40,0000	33,9000	47,6914	−13,7914
0,0000	21,0000	45,6000	29,6742	15,9258
0,0000	28,0000	20,6000	36,3121	−15,7121
0,0000	8,0000	3,8000	17,3467	−13,5467
0,0000	4,0000	4,4000	13,5536	−9,1536
0,0000	9,9800	28,9800	19,2243	9,7557
0,0000	1,0000	1,0000	10,7088	−9,7088
0,0000	14,0000	1,5000	23,0363	−21,5363
0,0000	1,0000	49,2000	10,7088	38,4912
0,0000	57,4800	55,0000	64,2672	−9,2672
0,0000	5,8000	35,2000	15,2605	19,9395
0,0000	1,2000	8,8000	10,8985	−2,0985
0,0000	0,9800	13,1000	10,6898	2,4102
0,0000	14,1000	8,3000	23,1312	−14,8312
0,0000	9,4800	50,2000	18,7501	31,4499
0,0000	24,0000	22,4000	32,5190	−10,1190
0,0000	24,0000	22,4000	32,5190	−10,11903
1,0000	42,0000	36,2000	38,5034	−2,3034
1,0000	24,0000	13,6000	21,4345	−7,8345
1,0000	25,0000	26,9000	22,3828	4,5172
1,0000	26,0000	15,8000	23,3311	−7,5311
0,0000	40,0000	40,2000	47,6914	−7,4914

```
0,0000   6,0000   8,8000   15,4502  -6,6502
0,0000   100,0000          123,2000          104,5877          18,6123
0,0000   21,0000  45,2000  29,6742  15,5258
0,0000   84,0000  66,4000  89,4153  -23,0153
0,0000   10,0000  6,0000   19,2432  -13,2432
0,0000   110,0000          107,6000          114,0704          -6,4704
0,0000   60,0000  33,0000  66,6568  -33,6568
0,0000   40,0000  63,0500  47,6914  15,3586
0,0000   21,0000  21,0000  29,6742  -8,6742

Residuals Mean: 11,6418

Degrees of freedom: 102
Number of data points: 2
Number of estimated paramaters: 3
```

Note que o número de variáveis independentes é menor do que o número de aplicações cadastradas, indicando que nem todas aplicações foram transformadas em variáveis *Dummy*.

Neste relatório possuímos diversas informações sobre a regressão, como as melhores estimativas para os coeficientes, parâmetros estimados, resíduos, valores calculados, etc. No momento, desejamos nos ater a análise dos resíduos da regressão, pois com isto, podemos encontrar possíveis pontos discrepantes. Resíduos proporcionalmente maiores do que a média indicam que uma entrada pode representar uma exceção, um ponto discrepante.

Analisando o relatório, podemos encontrar alguns possíveis casos, onde o resíduo é relativamente maior do que a média do conjunto, como pode ser visto na Tabela 6.3.

Requisito	**Incidents**	**Desenvolvimento (Delphi)**	**Esforço Real**	**Esforço Calculado**	**Resíduo**
Req 1	0	25,03	57,4	33,4958	23,9042
Req 8	1	21,08	48,1	18,6656	29,4344
Req 18	1	4,58	41,5	3,0191	38,4809
Req 69	0	25,4	70,4	33,8466	36,5534
Req 72	0	80	55	85,6223	-30,6223
Req 73	0	28	163,4	36,3121	127,0879
Req 74	0	73	127	78,9843	48,0157
Req 85	0	1	49,2	10,7088	38,4912
Req 92	0	9,48	50,2	18,7501	31,4499
Req 106	0	60	33	66,6568	-33,6568

Tabela 6.3: Requisitos do "Projeto 1"com resíduo elevado, modelo "Coding and Unit Test"

Estes requisitos provavelmente são pontos discrepantes, e para que não influenciem o modelo de regressão, devemos descartá-los.

Uma vez que descartados os requisitos acima, podemos gerar o relatório de regressão novamente, como pode ser visto abaixo:

```
Unweighted Least Squares Minimisation
Linear Regression with intercept
y = c[0] + c[1]*x1 + c[2]*x2 +c[3]*x3 + . . .
```

```
        Best Estimates
c[0]    4,5477
c[1]    -8,1008
c[2]    0,9846

Correlation: x - y data
Linear Correlation Coefficient (R): 0,9121
Linear Correlation Coefficient Squared (R^2): 0,8319

x1      x2      y(expl) y(calc) residual
0,0000  7,3800  21,7000 11,8142 9,8858
0,0000  53,7200 61,6000 57,4419 4,1581
0,0000  10,0000 15,0000 14,3939 0,6061
0,0000  16,8300 29,2000 21,1189 8,0811
1,0000  3,1700  13,5000 -0,4319 13,9319
1,0000  3,6700  3,8000  0,0605  3,7395
1,0000  4,1700  2,5000  0,5528  1,9472
1,0000  18,6700 4,8000  14,8299 -10,0299
1,0000  5,6700  8,8000  2,0297  6,7703
1,0000  4,5800  6,0000  0,9565  5,0435
1,0000  18,6700 5,0000  14,8299 -9,8299
1,0000  18,6700 6,5000  14,8299 -8,3299
1,0000  7,5800  3,8000  3,9104  -0,1104
1,0000  3,0800  1,0000  -0,5205 1,5205
1,0000  16,5000 6,0000  12,6932 -6,6932
1,0000  19,6700 28,6000 15,8145 12,7855
1,0000  9,1700  2,0000  5,4759  -3,4759
1,0000  22,9200 9,8000  19,0146 -9,2146
1,0000  16,5000 20,9000 12,6932 8,2068
1,0000  32,0000 18,6000 27,9550 -9,3550
1,0000  33,8300 31,2000 29,7569 1,4431
1,0000  33,8300 7,8000  29,7569 -21,9569
1,0000  3,0800  1,0000  -0,5205 1,5205
1,0000  6,5800  1,0000  2,9257  -1,9257
1,0000  3,1700  1,0000  -0,4319 1,4319
1,0000  3,1700  1,0000  -0,4319 1,4319
1,0000  3,0800  0,2000  -0,5205 0,7205
1,0000  9,0800  17,2000 5,3873  11,8127
1,0000  9,0800  8,8000  5,3873  3,4127
1,0000  22,1700 30,9000 18,2761 12,6239
1,0000  3,1700  3,0000  -0,4319 3,4319
1,0000  11,0800 8,8000  7,3566  1,4434
1,0000  14,0000 4,0000  10,2317 -6,2317
1,0000  14,0000 5,0000  10,2317 -5,2317
1,0000  14,0000 5,0000  10,2317 -5,2317
1,0000  23,5000 30,2000 19,5856 10,6144
1,0000  14,0000 24,4000 10,2317 14,1683
1,0000  14,0000 7,8000  10,2317 -2,4317
1,0000  14,0000 3,0000  10,2317 -7,2317
1,0000  14,0000 8,0000  10,2317 -2,2317
0,0000  30,0000 35,0000 34,0865 0,9135
0,0000  62,0000 74,6300 65,5946 9,0354
0,0000  40,0000 59,5000 43,9328 15,5672
0,0000  20,0000 20,0000 24,2402 -4,2402
0,0000  32,0000 47,8000 36,0558 11,7442
0,0000  32,0000 32,0000 36,0558 -4,0558
0,0000  8,0000  8,8000  12,4247 -3,6247
0,0000  16,0000 13,6000 20,3017 -6,7017
0,0000  8,0000  7,8000  12,4247 -4,6247
0,0000  40,0000 52,1000 43,9328 8,1672
0,0000  3,5000  3,5000  7,9939  -4,4939
0,0000  8,0000  6,8000  12,4247 -5,6247
0,0000  10,0000 10,0000 14,3939 -4,3939
0,0000  3,0000  2,0000  7,5015  -5,5015
0,0000  6,0000  4,0000  10,4554 -6,4554
0,0000  12,0000 5,8000  16,3632 -10,5632
0,0000  10,0000 10,4000 14,3939 -3,9939
0,0000  8,0000  21,3000 12,4247 8,8753
0,0000  8,0000  8,8000  12,4247 -3,6247
0,0000  6,0000  3,0000  10,4554 -7,4554
0,0000  6,3000  8,8000  10,7508 -1,9508
0,0000  16,0000 16,8000 20,3017 -3,5017
```

```
0,0000   25,4000 41,2000 29,5572 11,6428
0,0000   16,0000 18,1000 20,3017 -2,2017
0,0000   25,4000 21,6000 29,5572 -7,9572
0,0000   10,0000 10,0000 14,3939 -4,3939
0,0000   16,0000 24,7000 20,3017 4,3983
0,0000   40,0000 33,9000 43,9328 -10,0328
0,0000   21,0000 45,6000 25,2248 20,3752
0,0000   28,0000 20,6000 32,1172 -11,5172
0,0000   8,0000  3,8000  12,4247 -8,6247
0,0000   4,0000  4,4000  8,4862  -4,0862
0,0000   9,9800  28,9800 14,3742 14,6058
0,0000   1,0000  1,0000  5,5323  -4,5323
0,0000   14,0000 1,5000  18,3325 -16,8325
0,0000   57,4800 55,0000 61,1441 -6,1441
0,0000   5,8000  35,2000 10,2585 24,9415
0,0000   1,2000  8,8000  5,7292  3,0708
0,0000   0,9800  13,1000 5,5126  7,5874
0,0000   14,1000 8,3000  18,4309 -10,1309
0,0000   24,0000 22,4000 28,1787 -5,7787
0,0000   24,0000 22,4000 28,1787 -5,7787
1,0000   42,0000 36,2000 37,8013 -1,6013
1,0000   24,0000 13,6000 20,0780 -6,4780
1,0000   25,0000 26,9000 21,0626 5,8374
1,0000   26,0000 15,8000 22,0472 -6,2472
0,0000   40,0000 40,2000 43,9328 -3,7328
0,0000   6,0000  8,8000  10,4554 -1,6554
0,0000   100,0000        123,2000        103,0105        20,1895
0,0000   21,0000 45,2000 25,2248 19,9752
0,0000   84,0000 66,4000 87,2564 -20,8564
0,0000   10,0000 6,0000  14,3939 -8,3939
0,0000   110,0000        107,6000        112,8568        -5,2568
0,0000   40,0000 63,0500 43,9328 19,1172
0,0000   21,0000 21,0000 25,2248 -4,2248

Residuals Mean: 7,3005

Degrees of freedom: 92
Number of data points: 2
Number of estimated paramaters: 3
```

Note que ainda podemos possuir alguns pontos discrepantes, mas iremos ignorá-los neste momento. Mesmo descartando todos os possíveis pontos discrepantes, nosso resíduo não deverá diminuir significativamente, e também, podemos perder precisão no modelo, pois menos requisitos estarão sendo analisados. Devemos ter cuidado, principalmente neste ponto, onde não possuímos uma base de requisitos vasta, pois descartar requisitos em demasia pode gerar uma perda significativa na precisão do modelo.

Note também que o "Coeficiente de Correlação Linear" aumentou significativamente, indicando que após a retirada dos possíveis pontos discrepantes, nossas variáveis independentes estão "explicando" mais de nossa variável dependente, conforme mencionamos anteriormente no trabalho.

Uma vez realizada a análise sobre o modelo "Coding and Unit Test", devemos realizar o mesmo procedimento sobre o modelo "SIT Execution". Inicialmente desabilitamos alguns requisitos que possuíam dados incompletos ou estavam incorretos, logo após, geramos o primeiro relatório de regressão, que pode ser visto abaixo:

```
Unweighted Least Squares Minimisation
Linear Regression with intercept
y = c[0] + c[1]*x1 + c[2]*x2 +c[3]*x3 + . . .

        Best Estimates
c[0]    5,6766
c[1]    -5,9890
c[2]    0,4201
c[3]    0,2218

Correlation: x - y data
Linear Correlation Coefficient (R): 0,8151
Linear Correlation Coefficient Squared (R^2): 0,6644

x1      x2      x3      y(expl) y(calc) residual
0,0000  26,8000 43,0000 44,3000 26,4720 17,8280
0,0000  14,0000 9,0000  7,5000  13,5542 -6,0542
0,0000  7,0000  14,0000 10,0000 11,7223 -1,7223
0,0000  5,8000  3,0000  7,5000  8,7786  -1,2786
0,0000  17,5000 6,0000  7,0000  14,3593 -7,3593
1,0000  3,5000  41,0000 13,6000 10,2507 3,3493
1,0000  9,9200  76,0000 12,3000 20,7099 -8,4099
1,0000  1,1700  4,0000  3,0000  1,0663  1,9337
1,0000  2,3300  9,0000  2,5000  2,6625  -0,1625
1,0000  2,9200  10,0000 5,0000  3,1321  1,8679
1,0000  2,3300  10,0000 2,5000  2,8843  -0,3843
1,0000  4,6700  16,0000 2,0000  5,1980  -3,1980
1,0000  3,5000  21,0000 2,0000  5,8153  -3,8153
1,0000  4,6700  7,0000  2,0000  3,2020  -1,2020
1,0000  0,5800  5,0000  0,8000  1,0402  -0,2402
1,0000  0,5800  4,0000  1,0000  0,8184  0,1816
1,0000  3,5000  3,0000  5,8000  1,8234  3,9766
1,0000  1,7500  2,0000  2,0000  0,8664  1,1336
1,0000  2,3300  7,0000  2,0000  2,2189  -0,2189
1,0000  4,6700  6,0000  3,5000  2,9802  0,5198
1,0000  4,6700  4,0000  3,5000  2,5367  0,9633
1,0000  4,0800  4,0000  2,0000  2,2888  -0,2888
1,0000  1,7500  4,0000  2,5000  1,3100  1,1900
1,0000  1,7500  4,0000  4,5000  1,3100  3,1900
1,0000  1,7500  7,0000  0,5000  1,9753  -1,4753
1,0000  1,1700  8,0000  0,5000  1,9534  -1,4534
1,0000  1,7500  4,0000  0,5000  1,3100  -0,8100
1,0000  0,5800  3,0000  1,8000  0,5967  1,2033
1,0000  2,3300  9,0000  2,0000  2,6625  -0,6625
1,0000  2,3300  7,0000  3,5000  2,2189  1,2811
1,0000  5,8300  7,0000  5,0000  3,6894  1,3106
1,0000  3,5000  2,0000  3,5000  1,6016  1,8984
1,0000  3,5000  5,0000  4,5000  2,2669  2,2331
1,0000  4,6700  4,0000  4,0000  2,5367  1,4633
1,0000  4,6700  4,0000  4,0000  2,5367  1,4633
1,0000  4,6700  4,0000  3,5000  2,5367  0,9633
1,0000  7,0000  5,0000  6,0000  3,7373  2,2627
1,0000  5,8300  10,0000 5,0000  4,3547  0,6453
1,0000  3,5000  7,0000  2,0000  2,7105  -0,7105
1,0000  4,6700  4,0000  3,0000  2,5367  0,4633
1,0000  4,6700  5,0000  3,0000  2,7585  0,2415
0,0000  14,0000 5,0000  16,5000 12,6671 3,8329
0,0000  14,0000 4,0000  9,3000  12,4454 -3,1454
0,0000  23,3300 16,0000 12,4000 19,0263 -6,6263
0,0000  14,0000 3,0000  5,5000  12,2236 -6,7236
0,0000  8,1700  7,0000  5,0000  10,6614 -5,6614
0,0000  5,8300  13,0000 8,8000  11,0089 -2,2089
0,0000  5,8000  5,0000  3,5000  9,2222  -5,7222
0,0000  3,5000  8,0000  3,0000  8,9212  -5,9212
0,0000  5,8300  7,0000  16,1000 9,6783  6,4217
0,0000  2,3300  37,0000 53,7000 14,8611 38,8389
0,0000  4,6700  9,0000  21,8000 9,6345  12,1655
0,0000  4,6700  27,0000 21,3000 13,6264 7,6736
0,0000  9,3300  3,0000  15,0000 10,2616 4,7384
0,0000  2,3300  3,0000  11,5000 7,3208  4,1792
0,0000  4,6700  7,0000  9,2500  9,1910  0,0590
0,0000  4,6700  11,0000 14,5000 10,0781 4,4219
0,0000  9,3300  33,0000 19,2700 16,9148 2,3552
```

```
0,0000   4,6700   10,0000  29,0000  9,8563   19,1437
0,0000   3,5000   7,0000   14,5000  8,6994   5,8006
0,0000   3,5000   3,0000   11,5000  7,8124   3,6876
0,0000   2,3300   5,0000   4,0000   7,7644   -3,7644
0,0000   7,0000   7,0000   12,5000  10,1698  2,3302
0,0000   4,6700   6,0000   8,0200   8,9692   -0,9492
0,0000   2,3300   4,0000   9,3000   7,5426   1,7574
0,0000   9,3300   6,0000   6,8000   10,9269  -4,1269
0,0000   7,0000   5,0000   4,8000   9,7263   -4,9263
0,0000   4,6700   2,0000   2,3000   8,0821   -5,7821
0,0000   4,7000   1,0000   2,8000   7,8729   -5,0729
0,0000   4,6700   2,0000   3,0000   8,0821   -5,0821
0,0000   35,0000  13,0000  33,2000  23,2638  9,9362
0,0000   7,0000   4,0000   3,5000   9,5045   -6,0045
0,0000   7,0000   5,0000   4,5000   9,7263   -5,2263
0,0000   4,6700   2,0000   3,0000   8,0821   -5,0821
0,0000   9,3300   16,0000  7,5000   13,1447  -5,6447
0,0000   2,3300   9,0000   2,5000   8,6514   -6,1514
0,0000   6,6500   16,0000  9,5000   12,0188  -2,5188
0,0000   10,0000  16,0000  23,8000  13,4261  10,3739
0,0000   12,0000  12,0000  5,2500   13,3793  -8,1293
0,0000   11,0000  30,0000  10,7500  16,9511  -6,2011
0,0000   7,0000   21,0000  27,9000  13,2747  14,6253
0,0000   23,0000  110,0000          46,1000  39,7342  6,3658
0,0000   9,0000   10,0000  9,0000   11,6754  -2,6754
0,0000   50,0000  172,0000          58,3000  64,8273  -6,5273
0,0000   8,0000   18,0000  8,0000   13,0295  -5,0295
0,0000   3,0000   4,0000   3,3000   7,8241   -4,5241
0,0000   6,0000   28,0000  6,0000   14,4069  -8,4069
0,0000   42,0000  37,0000  31,6000  31,5271  0,0729
0,0000   39,0000  40,0000  47,7000  30,9321  16,7679
0,0000   4,0000   6,0000   3,0000   8,6877   -5,6877
0,0000   3,0000   6,0000   4,0000   8,2676   -4,2676
1,0000   20,9800  69,0000  11,0000  23,8040  -12,8040
1,0000   4,9800   17,0000  6,0000   5,5500   0,4500
1,0000   5,0000   7,0000   6,0000   3,3407   2,6593
1,0000   14,0000  11,0000  7,0000   8,0088   -1,0088
0,0000   9,9800   27,0000  13,6000  15,8572  -2,2572
0,0000   5,0000   24,0000  5,4000   13,0997  -7,6997
0,0000   12,0000  5,0000   13,5000  11,8269  1,6731
0,0000   15,5000  10,0000  8,6000   14,4062  -5,8062
0,0000   20,0000  22,0000  18,8000  18,9580  -0,1580
0,0000   10,0000  9,0000   5,0000   11,8737  -6,8737
0,0000   34,9800  88,0000  36,7500  39,8883  -3,1383
0,0000   9,0000   14,0000  10,2500  12,5625  -2,3125
0,0000   13,9800  14,0000  17,2500  14,6547  2,5953
0,0000   25,0000  65,0000  25,4000  30,5947  -5,1947

Residuals Mean: 4,4664

Degrees of freedom: 101
Number of data points: 3
Number of estimated paramaters: 4
```

Analisando o relatório, podemos encontrar alguns possíveis casos de pontos discrepantes, como pode ser visto na Tabela 6.4.

Devemos descartar estes requisitos, e gerar o relatório de regressão novamente. Note que anteriormente, inabilitamos requisitos somente para o modelo "Coding and Unit Test". Devemos realizar o mesmo trabalho sobre o modelo "SIT Execution" a partir da lista de requisitos completa.

O relatório de regressão gerado pode ser visto abaixo:

Requisito	Incidents	Teste (Delphi)	Cenários	Esforço Real	Esforço Calculado	Resíduo
Req 1	0	26,8	43	44,3	26,472	17,828
Req 53	0	2,33	37	53,7	14,8611	38,8389
Req 54	0	4,67	9	21,8	9,6345	12,1655
Req 61	0	4,67	10	29	9,8563	19,1437
Req 74	0	35	13	33,2	23,2638	9,9362
Req 81	0	10	16	23,8	13,4261	10,3739
Req 84	0	7	21	27,9	13,2747	14,6253
Req 92	0	39	40	47,7	30,9321	16,7679
Req 95	1	20,98	69	11	23,804	-12,804

Tabela 6.4: Requisitos do "Projeto 1"com erro elevado, modelo "SIT Execution"

```
Unweighted Least Squares Minimisation
Linear Regression with intercept
y = c[0] + c[1]*x1 + c[2]*x2 +c[3]*x3 + . . .

        Best Estimates
c[0]    3,7338
c[1]    -3,5353
c[2]    0,3699
c[3]    0,2225

Correlation: x - y data
Linear Correlation Coefficient (R): 0,9197
Linear Correlation Coefficient Squared (R^2): 0,8459

x1      x2      x3      y(expl) y(calc) residual
0,0000  14,0000 9,0000  7,5000  10,9153 -3,4153
0,0000  7,0000  14,0000 10,0000 9,4382  0,5618
0,0000  5,8000  3,0000  7,5000  6,5469  0,9531
0,0000  17,5000 6,0000  7,0000  11,5426 -4,5426
1,0000  3,5000  41,0000 13,6000 10,6152 2,9848
1,0000  9,9200  76,0000 12,3000 20,7773 -8,4773
1,0000  1,1700  4,0000  3,0000  1,5213  1,4787
1,0000  2,3300  9,0000  2,5000  3,0628  -0,5628
1,0000  2,9200  10,0000 5,0000  3,5036  1,4964
1,0000  2,3300  10,0000 2,5000  3,2853  -0,7853
1,0000  4,6700  16,0000 2,0000  5,4859  -3,4859
1,0000  3,5000  21,0000 2,0000  6,1655  -4,1655
1,0000  4,6700  7,0000  2,0000  3,4835  -1,4835
1,0000  0,5800  5,0000  0,8000  1,5255  -0,7255
1,0000  0,5800  4,0000  1,0000  1,3030  -0,3030
1,0000  3,5000  3,0000  5,8000  2,1607  3,6393
1,0000  1,7500  2,0000  2,0000  1,2908  0,7092
1,0000  2,3300  7,0000  2,0000  2,6178  -0,6178
1,0000  4,6700  6,0000  3,5000  3,2610  0,2390
1,0000  4,6700  4,0000  3,5000  2,8160  0,6840
1,0000  4,0800  4,0000  2,0000  2,5978  -0,5978
1,0000  1,7500  4,0000  2,5000  1,7358  0,7642
1,0000  1,7500  4,0000  4,5000  1,7358  2,7642
1,0000  1,7500  7,0000  0,5000  2,4033  -1,9033
1,0000  1,1700  8,0000  0,5000  2,4112  -1,9112
1,0000  1,7500  4,0000  0,5000  1,7358  -1,2358
1,0000  0,5800  3,0000  1,8000  1,0805  0,7195
1,0000  2,3300  9,0000  2,0000  3,0628  -1,0628
1,0000  2,3300  7,0000  3,5000  2,6178  0,8822
1,0000  5,8300  7,0000  5,0000  3,9126  1,0874
1,0000  3,5000  2,0000  3,5000  1,9382  1,5618
1,0000  3,5000  5,0000  4,5000  2,6057  1,8943
1,0000  4,6700  4,0000  4,0000  2,8160  1,1840
1,0000  4,6700  4,0000  4,0000  2,8160  1,1840
1,0000  4,6700  4,0000  3,5000  2,8160  0,6840
1,0000  7,0000  5,0000  6,0000  3,9005  2,0995
1,0000  5,8300  10,0000 5,0000  4,5801  0,4199
1,0000  3,5000  7,0000  2,0000  3,0507  -1,0507
1,0000  4,6700  4,0000  3,0000  2,8160  0,1840
1,0000  4,6700  5,0000  3,0000  3,0385  -0,0385
0,0000  14,0000 5,0000  16,5000 10,0253 6,4747
```

```
0,0000   14,0000 4,0000   9,3000   9,8029   -0,5029
0,0000   23,3300 16,0000  12,4000  15,9242  -3,5242
0,0000   14,0000 3,0000   5,5000   9,5804   -4,0804
0,0000   8,1700  7,0000   5,0000   8,3136   -3,3136
0,0000   5,8300  13,0000  8,8000   8,7829   0,0171
0,0000   5,8000  5,0000   3,5000   6,9919   -3,4919
0,0000   3,5000  8,0000   3,0000   6,8085   -3,8085
0,0000   5,8300  7,0000   16,1000  7,4480   8,6520
0,0000   4,6700  27,0000  21,3000  11,4686  9,8314
0,0000   9,3300  3,0000   15,0000  7,8528   7,1472
0,0000   2,3300  3,0000   11,5000  5,2632   6,2368
0,0000   4,6700  7,0000   9,2500   7,0188   2,2312
0,0000   4,6700  11,0000  14,5000  7,9088   6,5912
0,0000   9,3300  33,0000  19,2700  14,5274  4,7426
0,0000   3,5000  7,0000   14,5000  6,5860   7,9140
0,0000   3,5000  3,0000   11,5000  5,6961   5,8039
0,0000   2,3300  5,0000   4,0000   5,7082   -1,7082
0,0000   7,0000  7,0000   12,5000  7,8808   4,6192
0,0000   4,6700  6,0000   8,0200   6,7963   1,2237
0,0000   2,3300  4,0000   9,3000   5,4857   3,8143
0,0000   9,3300  6,0000   6,8000   8,5202   -1,7202
0,0000   7,0000  5,0000   4,8000   7,4358   -2,6358
0,0000   4,6700  2,0000   2,3000   5,9064   -3,6064
0,0000   4,7000  1,0000   2,8000   5,6950   -2,8950
0,0000   4,6700  2,0000   3,0000   5,9064   -2,9064
0,0000   7,0000  4,0000   3,5000   7,2133   -3,7133
0,0000   7,0000  5,0000   4,5000   7,4358   -2,9358
0,0000   4,6700  2,0000   3,0000   5,9064   -2,9064
0,0000   9,3300  16,0000  7,5000   10,7451  -3,2451
0,0000   2,3300  9,0000   2,5000   6,5982   -4,0982
0,0000   6,6500  16,0000  9,5000   9,7537   -0,2537
0,0000   12,0000 12,0000  5,2500   10,8429  -5,5929
0,0000   11,0000 30,0000  10,7500  14,4777  -3,7277
0,0000   23,0000 110,0000          46,1000  36,7159  9,3841
0,0000   9,0000  10,0000  9,0000   9,2881   -0,2881
0,0000   50,0000 172,0000          58,3000  60,4984  -2,1984
0,0000   8,0000  18,0000  8,0000   10,6981  -2,6981
0,0000   3,0000  4,0000   3,3000   5,7336   -2,4336
0,0000   6,0000  28,0000  6,0000   12,1831  -6,1831
0,0000   42,0000 37,0000  31,6000  27,5031  4,0969
0,0000   4,0000  6,0000   3,0000   6,5485   -3,5485
0,0000   3,0000  6,0000   4,0000   6,1786   -2,1786
1,0000   4,9800  17,0000  6,0000   5,8230   0,1770
1,0000   5,0000  7,0000   6,0000   3,6056   2,3944
1,0000   14,0000 11,0000  7,0000   7,8249   -0,8249
0,0000   9,9800  27,0000  13,6000  13,4329  0,1671
0,0000   5,0000  24,0000  5,4000   10,9232  -5,5232
0,0000   12,0000 5,0000   13,5000  9,2855   4,2145
0,0000   15,5000 10,0000  8,6000   11,6927  -3,0927
0,0000   20,0000 22,0000  18,8000  16,0272  2,7728
0,0000   10,0000 9,0000   5,0000   9,4356   -4,4356
0,0000   34,9800 88,0000  36,7500  36,2530  0,4970
0,0000   9,0000  14,0000  10,2500  10,1781  0,0719
0,0000   13,9800 14,0000  17,2500  12,0203  5,2297
0,0000   25,0000 65,0000  25,4000  27,4439  -2,0439

Residuals Mean: 2,7600

Degrees of freedom: 92
Number of data points: 3
Number of estimated paramaters: 4
```

Note novamente que ainda podemos possuir alguns pontos discrepantes, mas também iremos ignorá-los neste momento pelo mesmo motivo mencionado anteriormente.

Note também que novamente o "Coeficiente de Correlação Linear" aumentou significativamente, indicando que após a retirada dos possíveis pontos discrepantes, nossas

variáveis independentes estão "explicando" mais de nossa variável dependente.

Uma vez que temos modelos de regressão satisfatórios, podemos adicionar novos projetos ao sistema para verificar como as estimativas dos coeficientes da regressão irão se comportar com novos requisitos.

Como mencionado anteriormente, estamos utilizando dados reais de projetos. O segundo projeto a ser adicionado será denominado "Projeto 2" e irá conter os requisitos apresentados na Tabela 6.5

Requisito	**Aplicação**	**Desenvolvimento (Delphi)**	**Teste (Delphi)**	**Cenários de Teste**
Req 109	App 10	17	32	57
Req 110	App 10	35	13	18
Req 111	App 10	50	12	18
Req 112	App 10	8	28	35
Req 113	App 1	14	10	11
Req 114	App 1	70	20	24
Req 115	App 6	57,6	7	13
Req 116	App 6	30	16	15
Req 117	App 13	50	14	6
Req 118	App 13	60	8	15
Req 119	App 4	20,7	10	19
Req 120	App 4	40	9	15
Req 121	App 4	8	10	9
Req 122	App 5	28	11	8
Req 123	App 5	7	3	4
Req 124	App 5	20	5	8
Req 125	App 5	10	4	5
Req 126	App 5	28	6	5

Tabela 6.5: Requisitos populados no "Projeto 2"

Criado o novo projeto, devemos adicionar todos os seus requisitos ao sistema, conforme descrito no Apêndice B, pela funcionalidade "Add Requirement". Neste ponto, ainda não devemos informar quais os atributos de cada requisito, somente seu nome e suas tarefas. Para todos os requisitos, selecionaremos ambas tarefas, "Coding and Unit Test" e "SIT Exection" para que possamos estimar seu esforço durante a realização do projeto. Neste ponto, cada requisito e suas tarefas são adicionados ao sistema.

Uma vez que todos os requisitos são adicionados, podemos iniciar a primeira fase do projeto, neste caso "Development', conforme descrito no Apêndice B, pela funcionalidade "Start Phase". Quando requisitamos o início de uma fase, devemos informar os atributos refentes aos requisitos que possuem tarefas cadastradas na mesma. A partir destes valores, o sistema irá calcular sua estimativa, baseado nos coeficientes gerados pela regressão. Abaixo pode ser vistas as estimativas geradas pelo sistema:

```
Coding and Unit Test Estimates:
Requirement                    Estimate (Hours)
Req 109                        22,3975
Req 110                        39,7545
```

```
Req 111                         54,2187
Req 112                         13,7190
Req 113                         19,5046
Req 114                         73,5043
Req 115                         61,5472
Req 116                         34,9331
Req 117                         54,2187
Req 118                         63,8615
Req 119                         25,9653
Req 120                         44,5759
Req 121                         13,7190
Req 122                         28,4478
Req 123                         8,1979
Req 124                         20,7336
Req 125                         11,0908
Req 126                         28,4478
```

Após a execução de todas tarefas da fase, podemos finaliza-la pela funcionalidade "Finish Phase", informando o esforço real necessário para execução de cada tarefa. Estes valores podem ser vistos na Tabela 6.6

Requisito	**Tarefa**	**Esforço**
Req 109	Coding and Unit Test	17
Req 110	Coding and Unit Test	35
Req 111	Coding and Unit Test	16,8
Req 112	Coding and Unit Test	15,77
Req 113	Coding and Unit Test	14
Req 114	Coding and Unit Test	71
Req 115	Coding and Unit Test	57,6
Req 116	Coding and Unit Test	45,4
Req 117	Coding and Unit Test	48
Req 118	Coding and Unit Test	72,2
Req 119	Coding and Unit Test	20,7
Req 120	Coding and Unit Test	41,8
Req 121	Coding and Unit Test	17,3
Req 122	Coding and Unit Test	28
Req 123	Coding and Unit Test	7
Req 124	Coding and Unit Test	20
Req 125	Coding and Unit Test	0
Req 126	Coding and Unit Test	37,5

Tabela 6.6: Esforço real para tarefa "Coding and Unit Test" no "Projeto 2"

Note que possuímos um valor zerado. Como estamos utilizando dados reais de projetos, provavelmente este valores está incorreto, e posteriormente deverá ser descartado para que não influencie as equações de regressão.

Neste ponto, podemos gerar um relatório do projeto, a fim de analisar seu andamento e resultados. O relatório gerado pode ser visto abaixo:

```
Project Report

Project: Projeto 2

Phase    Started Finished
Development      1        1
Testing 0        0

Phase    Task    Requirement             Estimated        Actual   Error(%)
Development      Coding and Unit Te      Req 109 22.3975 17.0000 31,7500
Testing          SIT Execution           Req 109 null     null     null
```

```
Development     Coding and Unit Te      Req 110 39.7545 35.0000 13,5843
Testing         SIT Execution           Req 110 null    null    null
Development     Coding and Unit Te      Req 111 54.2187 16.8000 222,7304
Testing         SIT Execution           Req 111 null    null    null
Development     Coding and Unit Te      Req 112 13.7190 15.7700 -13,0057
Testing         SIT Execution           Req 112 null    null    null
Development     Coding and Unit Te      Req 113 19.5046 14.0000 39,3186
Testing         SIT Execution           Req 113 null    null    null
Development     Coding and Unit Te      Req 114 73.5043 71.0000 3,5272
Testing         SIT Execution           Req 114 null    null    null
Development     Coding and Unit Te      Req 115 61.5472 57.6000 6,8528
Testing         SIT Execution           Req 115 null    null    null
Development     Coding and Unit Te      Req 116 34.9331 45.4000 -23,0548
Testing         SIT Execution           Req 116 null    null    null
Development     Coding and Unit Te      Req 117 54.2187 48.0000 12,9556
Testing         SIT Execution           Req 117 null    null    null
Development     Coding and Unit Te      Req 118 63.8615 72.2000 -11,5492
Testing         SIT Execution           Req 118 null    null    null
Development     Coding and Unit Te      Req 119 25.9653 20.7000 25,4362
Testing         SIT Execution           Req 119 null    null    null
Development     Coding and Unit Te      Req 120 44.5759 41.8000 6,6409
Testing         SIT Execution           Req 120 null    null    null
Development     Coding and Unit Te      Req 121 13.7190 17.3000 -20,6994
Testing         SIT Execution           Req 121 null    null    null
Development     Coding and Unit Te      Req 122 28.4478 28.0000 1,5993
Testing     SIT Execution       Req 122 null    null    null
Development     Coding and Unit Te      Req 123 8.1979  7.0000  17,1129
Testing         SIT Execution           Req 123 null    null    null
Development     Coding and Unit Te      Req 124 20.7336 20.0000 3,6680
Testing         SIT Execution           Req 124 null    null    null
Development     Coding and Unit Te      Req 125 11.0908 0.0000  Infinity
Testing         SIT Execution           Req 125 null    null    null
Development     Coding and Unit Te      Req 126 28.4478 37.5000 -24,1392
Testing         SIT Execution           Req 126 null    null    null
```

Este relatório nos traz algumas informações interessantes, como quais as fases que possui, quais foram iniciadas e finalizadas, suas tarefas, juntamente com o esforço estimado e real para cada uma delas e seu erro percentual. Note que possuímos alguns valores nulos, indicando que estes dados ainda não estão disponíveis.

Uma análise interessante que podemos realizar sobre este relatório é sobre os erros que a estimativa de esforço teve sobre o esforço real. Mencionamos anteriormente neste trabalho que uma boa abordagem de estimativa deve prover estimativas que estão entre 25% dos resultados reais em 75% do tempo [CDS86]. Baseados nesta afirmação, podemos realizar uma analise percentil sobre os erros percentuais absolutos gerados pelas estimativas.

A partir desta análise, descobrimos que o 75° percentil dos erros percentuais absolutos gerados neste projeto na tarefa "Coding and Unit Test" é 24,1392%, indicando que em 75% dos casos, o erro é menor do que 24,1392%, o que pela definição dada anteriormente, representa uma boa abordagem de estimativa. Na Figura 6.1 podemos ver os erros percentuais absolutos ordenados de forma ascendente.

Note que no relatório acima, o erro refente ao requisito "Req 111" é muito maior do que os demais, indicando que provavelmente este caso seja uma exceção, um ponto

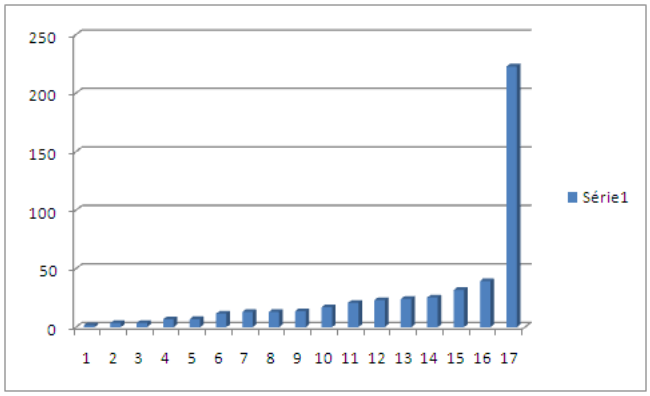


Figura 6.1: Erros percentuais absolutos da tarefa “Coding and Unit Test” no “Projeto 2”

discrepante, e que futuramente deverá ser descartado para que não influencie as equações de regressão.

Outra análise que podemos realizar, é sobre os resíduos do projeto. Podemos ver abaixo na Tabela 6.7, os resíduos gerados por cada requisito na tarefa “Coding and Unit Test”.

Requisito	**Estimado (Horas)**	**Real (Horas)**	**Resíduo (Horas)**
Req 109	28,2535	38,38	-10,1265
Req 110	12,5478	15,9	-3,3522
Req 111	12,1778	10,18	1,9978
Req 112	21,8791	18,4	3,4791
Req 113	9,8805	11,5	-1,6195
Req 114	16,4722	21,6	-5,1278
Req 115	9,2157	8	1,2157
Req 116	12,9901	11	1,9901
Req 117	10,2478	8,3	1,9478
Req 118	10,0306	7	3,0306
Req 119	11,6604	12,3	-0,6396
Req 120	10,4005	10,4	0,0005
Req 121	9,4356	10,8	-1,3644
Req 122	9,583	10	-0,417
Req 123	5,7336	4	1,7336
Req 124	7,3634	3	4,3634
Req 125	6,326	4	2,326
Req 126	7,0659	6	1,0659

Tabela 6.7: Resíduos para tarefa “Coding and Unit Test” no “Projeto 2”

Um dado interessante que obtivemos nesta análise foi resíduo total do projeto, 53,7672 horas, representando 9,5151% do esforço total do projeto, abaixo da média geral de erros percentuais absolutos de 24,1392% apresentanda anteriormente. Isto pode ser explicado pela Lei dos Grandes Números, citado anteriormente no trabalho, que nos diz que o erro absoluto do projeto tende a ser menor do que os erros absolutos individuais de cada estimativa, pois o erro negativo de uma estimativa anula o erro positivo de outro, e vice-versa.

Isto nos diz que se o projeto for analisado como um todo, nosso erro passa a ser 9,5151%, muito abaixo do que o esperado por uma boa abordagem de estimativa, como definido anteriormente.

Uma vez que finalizamos a fase "Development", podemos iniciar uma nova fase, neste caso, a fase "Testing". Novamente devemos informar os atributos refentes aos requisitos que possuem tarefas cadastradas nesta fase, para que o sistema possa gerar suas estimativas.

Abaixo pode ser vistas as estimativas geradas pelo sistema:

```
SIT Execution Estimates:
Requirement                    Estimate (Hours)
Req 109                        28,2535
Req 110                        12,5478
Req 111                        12,1778
Req 112                        21,8791
Req 113                        9,8805
Req 114                        16,4722
Req 115                        9,2157
Req 116                        12,9901
Req 117                        10,2478
Req 118                        10,0306
Req 119                        11,6604
Req 120                        10,4005
Req 121                        9,4356
Req 122                        9,5830
Req 123                        5,7336
Req 124                        7,3634
Req 125                        6,3260
Req 126                        7,0659
```

Novamente, após a execução de todas tarefas da fase, podemos finaliza-la e inserir o esforço real necessário a cada uma. Estes valores podem ser vistos na Tabela 6.8

Requisito	**Tarefa**	**Esforço**
Req 109	SIT Execution	38,38
Req 110	SIT Execution	15,9
Req 111	SIT Execution	10,18
Req 112	SIT Execution	18,4
Req 113	SIT Execution	11,5
Req 114	SIT Execution	21,6
Req 115	SIT Execution	8
Req 116	SIT Execution	11
Req 117	SIT Execution	8,3
Req 118	SIT Execution	7
Req 119	SIT Execution	12,3
Req 120	SIT Execution	10,4
Req 121	SIT Execution	10,8
Req 122	SIT Execution	10
Req 123	SIT Execution	4
Req 124	SIT Execution	3
Req 125	SIT Execution	4
Req 126	SIT Execution	6

Tabela 6.8: Esforço real para tarefa "SIT Execution" no "Projeto 2"

Após finalizarmos a fase "Testing", podemos novamente gerar o relatório do pro-

jeto, como pode ser visto abaixo:

Phase	Task	Requirement	Estimated		Actual	Error(%)
Development	Coding and Unit Te	Req 109	22.3975	17.0000	31,7500	
Testing	SIT Execution	Req 109	28.2535	38.3800	−26,3848	
Development	Coding and Unit Te	Req 110	39.7545	35.0000	13,5843	
Testing	SIT Execution	Req 110	12.5478	15.9000	−21,0830	
Development	Coding and Unit Te	Req 111	54.2187	16.8000	222,7304	
Testing	SIT Execution	Req 111	12.1778	10.1800	19,6248	
Development	Coding and Unit Te	Req 112	13.7190	15.7700	−13,0057	
Testing	SIT Execution	Req 112	21.8791	18.4000	18,9082	
Development	Coding and Unit Te	Req 113	19.5046	14.0000	39,3186	
Testing	SIT Execution	Req 113	9.8805	11.5000	−14,0826	
Development	Coding and Unit Te	Req 114	73.5043	71.0000	3,5272	
Testing	SIT Execution	Req 114	16.4722	21.6000	−23,7398	
Development	Coding and Unit Te	Req 115	61.5472	57.6000	6,8528	
Testing	SIT Execution	Req 115	9.2157	8.0000	15,1963	
Development	Coding and Unit Te	Req 116	34.9331	45.4000	−23,0548	
Testing	SIT Execution	Req 116	12.9901	11.0000	18,0918	
Development	Coding and Unit Te	Req 117	54.2187	48.0000	12,9556	
Testing	SIT Execution	Req 117	10.2478	8.3000	23,4675	
Development	Coding and Unit Te	Req 118	63.8615	72.2000	−11,5492	
Testing	SIT Execution	Req 118	10.0306	7.0000	43,2943	
Development	Coding and Unit Te	Req 119	25.9653	20.7000	25,4362	
Testing	SIT Execution	Req 119	11.6604	12.3000	−5,2000	
Development	Coding and Unit Te	Req 120	44.5759	41.8000	6,6409	
Testing	SIT Execution	Req 120	10.4005	10.4000	0,0048	
Development	Coding and Unit Te	Req 121	13.7190	17.3000	−20,6994	
Testing	SIT Execution	Req 121	9.4356	10.8000	−12,6333	
Development	Coding and Unit Te	Req 122	28.4478	28.0000	1,5993	
Testing	SIT Execution	Req 122	9.5830	10.0000	−4,1700	
Development	Coding and Unit Te	Req 123	8.1979	7.0000	17,1129	
Testing	SIT Execution	Req 123	5.7336	4.0000	43,3400	
Development	Coding and Unit Te	Req 124	20.7336	20.0000	3,6680	
Testing	SIT Execution	Req 124	7.3634	3.0000	145,4467	
Development	Coding and Unit Te	Req 125	11.0908	0.0000	Infinity	
Testing	SIT Execution	Req 125	6.3260	4.0000	58,1500	
Development	Coding and Unit Te	Req 126	28.4478	37.5000	−24,1392	
Testing	SIT Execution	Req 126	7.0659	6.0000	17,7650	

Realizamos novamente a análise percentil sobre os erros percentuais absolutos, mas desta vez sobre a tarefa “SIT Execution”, e obtivemos um erro de 25,7235% para o 75° percentil dos erros absolutos, levemente maior do que o desejado. Na Figura 6.2 podemos ver os erros percentuais absolutos ordenados de forma ascendente.

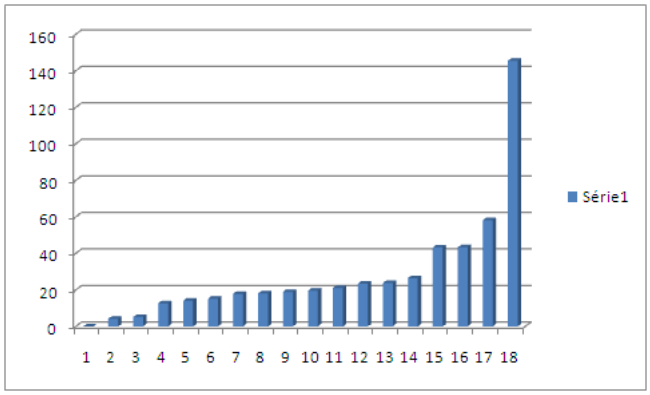


Figura 6.2: Erros percentuais absolutos da tarefa “SIT Execution” no “Projeto 2”

Note que neste caso, ao invés de um erro significativamente maior do que os demais, possuímos alguns erros acima de nosso 75° percentil. Estes erros podem ter sido gerados devido ao problema de se estimar valores além do intervalo de dados que foi utilizado para gerar os coeficientes de regressão, logo, nosso modelo não explica corretamente valores nesta faixa. A tendência é de que com o aumento da base de dados históricos, o intervalo de dados utilizado para gerar os coeficientes de regressão também aumente, aumentando assim, nossa faixa de estimativas.

Durante nossa análise sobre os modelos de regressão, o modelo "SIT Execution" nos pareceu "melhor explicado" do que o modelo "Coding and Unit Test", pois possuía resíduos menores e um coeficiente de correlação linear maior, mas mesmo assim, nos gerou um erro levemente maior.

Novamente faremos uma análise sobre os resíduos gerados nesta tarefa, a fim de identificar qual o erro absoluto do projeto como um todo. Os resíduos individuais podem ser vistos na Tabela 6.9.

Requisito	**Estimado (Horas)**	**Real (Horas)**	**Resíduo (Horas)**
Req 109	28,2535	38,38	-10,1265
Req 110	12,5478	15,9	-3,3522
Req 111	12,1778	10,18	1,9978
Req 112	21,8791	18,4	3,4791
Req 113	9,8805	11,5	-1,6195
Req 114	16,4722	21,6	-5,1278
Req 115	9,2157	8	1,2157
Req 116	12,9901	11	1,9901
Req 117	10,2478	8,3	1,9478
Req 118	10,0306	7	3,0306
Req 119	11,6604	12,3	-0,6396
Req 120	10,4005	10,4	0,0005
Req 121	9,4356	10,8	-1,3644
Req 122	9,583	10	-0,417
Req 123	5,7336	4	1,7336
Req 124	7,3634	3	4,3634
Req 125	6,326	4	2,326
Req 126	7,0659	6	1,0659

Tabela 6.9: Resíduos para tarefa "SIT Execution" no "Projeto 2"

Nesta análise obtivemos como resíduo total do projeto, 0,5035 horas, representando 0,2388% do esforço total do projeto, muito abaixo da média geral de erros percentuais absolutos de 25,7235% apresentanda anteriormente. Neste caso, a Lei dos Grandes Números nos favorece muito, pois quase todo resíduos foram anulados.

Se o projeto for analisado como um todo, nosso erro passa a ser apenas 0,2388%, um valor excelente e muito abaixo do que o esperado por uma boa abordagem de estimativa.

7 Conclusão

No início deste trabalho definimos que seu objetivo geral seria prover um mecanismo que possibilitasse a criação de um modelo de estimativas adaptável a diversos cenários, levando em consideração suas características particulares, a fim de obter um aumento na precisão final da estimativa.

A partir deste objetivo, podemos salientar duas necessidades primárias que nossa solução deveria prover, a grande capacidade de adaptação a diversos cenários, e uma alta confiabilidade em seus resultados.

A primeira necessidade pode ser satisfeita pelo método adotado para criação dos modelos, a regressão linear. Este método, como vimos anteriormente, possibilita a criação de modelos baseados nos mais diferentes conjuntos de dados, sejam eles dados quantitativos ou qualitativos. A grande vantagem do uso de regressão está na sua flexibilidade, podendo se adaptar a diversos cenários e podendo ser utilizado em praticamente todas as fases e para as mais diversas tarefas de um projeto. Isto possibilita a implantação da solução nos mais variados cenários, pois poderíamos utilizar uma vasta gama de atributos gerados por diversos processos de desenvolvimento, sem a necessidade de que o processo de desenvolvimento se adapte ao processo de estimativa, e sim o contrário.

A segunda necessidade pode ser atestada pelos resultados obtidos nos testes realizados, onde mesmo com dados escassos e uma pequena quantidade de atributos para serem analisados, obtivemos excelentes resultados nas estimativas geradas para um novo projeto. Em teoria, conforme adicionamos mais dados históricos ao sistema, a precisão do nosso modelo deverá aumentar, juntamente com seu intervalo de confiança, possibilitando assim, estimativas mais precisas sobre um conjunto de dados ainda maior.

Apesar dos resultados obtidos nos testes, não podemos afirmar que o mesmo irá ocorrer em diferentes situações, com diferentes dados e variáveis, pois cada modelo possui diferentes características que determinarão sua precisão, um modelo contendo variáveis altamente aleatórias provavelmente obterá um grau de precisão menor do que um modelo contendo variáveis mais previsíveis.

Apesar de todos os pontos positivos, devemos ter alguns cuidados no seu uso. Como mencionamos anteriormente, pontos discrepantes podem comprometer seriamente a qualidade da equação de regressão gerada pelo método. Tais pontos ocorrem com frequência em projetos reais, onde muitas vezes falhamos na análise dos requisitos, gerando assim, falhas nos dados inseridos no modelo. Devido a isto, um acompanhamento deve ser feito periodicamente para que não comprometamos a qualidade das estimativas geradas.

Outro ponto que devemos observar, principalmente na criação de novos modelos, é a inclusão de variáveis irrelevantes ou a omissão de variáveis relevantes no modelo. Tais problemas podem ocasionar a criação de um modelo que nunca obterá a precisão desejada, pois não poderá avaliar variáveis importantes à estimativa, ou irá adicionar erros de variáveis irrelevantes às mesmas. Por este motivo, devemos sempre realizar uma análise sobre quais atributos devemos utilizar na criação de um novo modelo.

Finalmente podemos concluir que, se bem utilizada, a solução apresentada pode nos fornecer estimativas muito confiáveis nos mais diferentes cenários, podendo assim, vir a ser uma ótima opção para solução deste problema que atualmente assola o setor.

8 Trabalhos futuros

Neste capítulo são apresentadas sugestões de trabalhos futuros motivados a partir da pesquisa apresentada neste trabalho.

8.1 Realização de testes com uma base de dados históricos maior

Como mencionamos anteriormente, a base de dados utilizada neste trabalho é relativamente pequena, abrangendo somente duas tarefas do projeto e contendo poucos requisitos. Para que um modelo de regressão gere estimativas confiáveis para uma grande faixa de valores, devemos possuir uma grande base de dados para geração das equações de regressão, principalmente se utilizarmos uma grande quantidade de dados qualitativos. Esta base de dados deve ser gerada pela contínua utilização da solução, pois isto garante um crescimento gradual, onde podemos analisar continuamente o modelo gerado, garantindo assim, que o mesmo não possua entradas incorretas ou pontos discrepantes.

Tendo em vista este fato, acreditamos que o contínuo uso da solução em um ambiente real nos trará mais informações sobre como cada modelo se comportará com uma grande quantidade de dados e nos auxiliará na procura de melhorias, tornando esta solução cada vez mais robusta e confiável.

8.2 Sistema para criação de modelos de estimativa baseados em regressão

Os conhecimentos produzidos a partir deste trabalho representam o ponto de partida para a concepção de um sistema que auxilie a criação de novos modelos de estimativa, facilitando assim esta difícil tarefa. Novos modelos poderiam ser criados para novas tarefas com facilidade, informando somente seus atributos e valores iniciais.

Com esta facilidade, poderíamos descrever processos mais completos mais rapida-

mente, pois a tarefa de criação e alteração de modelos seria simples. Com esta facilidade, poderíamos realizar análises mais detalhadas sobre os modelos, como a análise da relevância de variáveis adicionadas, pois poderíamos criar diversos modelos e realizar comparações entre os mesmos.

O sistema também poderia auxiliar na análise de resíduos de regressão e dados incorretos, automatizando este processo longo e custoso. Também poderia demonstrar graficamente informações sobre a regressão, como um gráfico de dispersão de pontos analisados e reta de regressão gerada para cada variável estimada.

Resumidamente, um sistema com tais características seria de grande auxílio a popularização desta solução, levando seu uso a um número maior de empresas que não possuam a capacidade ou disponibilidade para utilização da solução como está implementada atualmente.

Referências

[AG83] A. Albrecht and J. Gaffney. Software function, source lines of code, and development efford prediction: A software science validation. *IEEE Transactions on Software Engineering*, 1983.

[Agu03] Mauricio Aguiar. Pontos de função ou pontos por caso de uso? como estimar projetos orientados a objetos. *Developers Magazine*, 2003.

[Alb79] A. Albrecht. Measuring application development productivity. In *Proceedings of the Joint SHARE/GUIDE/IBM Application Development Symposium*, pages 83–92, Outubro 1979.

[Ans73] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27:12–21, 1973.

[BAC00] Barry W. Boehm, Chris Abts, and Sunita Chulani. Software development cost estimation approaches - A survey. *Ann. Software Eng*, 10:1–38, 2000.

[BGS84] Barry W. Boehm, T. E. Gray, and T. Seewaldt. Prototyping versus specifying: A multiproject experiment. *IEEE Transactions on Software Engineering*, 10:290–303, 1984.

[Boe81] Barry Boehm. *Software Engineering Economics*. Prentice-Hall, 1981.

[Boe87] B. Boehm. Industrial software metrics top ten list. *IEEE Software*, 4(5), Setembro 1987.

[Boe00] Barry Boehm. *Software Cost Estimation with COCOMO II*. Addison-Wesley, Reading, MA, 2000.

[BP88] B. Boehm and P. Papaccio. Understanding and controlling software costs. *IEEE Transactions on Software Engineering*, 14(10):1462–1477, Outubro 1988.

[Cam07] Fábio Martinho Campos. Métricas de software como ferramenta de apoio ao gerenciamento de projetos. *APInfo Website*, 2007.

[Car87] David N. Card. A software technology evaluation program. *Information And Software Technology*, 29:291–300, 1987.

[CDS86] Samuel D. Conte, Hubert E. Dunsmore, and Vincent Yun Shen. *Software Engineering Metrics and Models*. Benjamin / Cummings Publishing Company, Menlo Park, CA, 1986.

[Coh06] Mike Cohn. *Agile estimating and planning*. Prentice-Hall, 2006.

[CSB+86] B. Curtis, E. Soloway, R. Brooks, J. Black, K. Ehrlich, and H. Ramsey. Software psychology: The need for an interdisciplinary program. *Proceedings of the IEEE*, 74(8):1092–1106, August 1986.

[Cur81] Bill Curtis. Substantiating programmer variability. *Proceedings of the IEEE*, 69:846, 1981.

[DL85] Tom DeMarco and Tim Lister. Programmer performance and the effects of the workplace. In *ICSE*, pages 268–272, 1985.

[Gil88] Tom Gilb. *Principles of Software Engineering Management*. Addison-Wesley, 1988.

[Gla94] Robert L. Glass. Is field: Stress up, satisfaction down. Software Pratictioner, 1994.

[GW91] John Gaffney and Richard Werling. Estimating software size from counts of externals, a generalization of function points. *Software Productivity Consortium*, 1991.

[Her00] American Heritage. *The American Heritage Dictionary of the English Language*. Houghton Mifflin Company, Boston, 2000.

[HHA91] J. Hihn and H. Habib-Agahi. Cost estimation of software intensive projects: a survey of current practices. *International Conference on Software Engineering*, pages 276–287, 1991.

[Jen83] R. Jensen. An improved macrolevel software development resource estimation model. In *5th ISPA Conference*, pages 88–92, Abril 1983.

[Jon86] Capers Jones. *Programming Productivity*. McGraw-Hill, 1986.

[Jon91] Capers Jones. *Applied Software Measurement: Assuring Productivity and Quality*. McGraw-Hill, New York, NY, 1991.

[Jon98] Capers Jones. *Estimating Software Costs*. McGraw-Hill, New York, NY, 1998.

[Jor02] M. Jorgensen. *A Review of Studies on Expert Estimation of Software Development Effort*. 2002.

[Jor04] Magne Jorgensen. A review of studies on expert estimation of software development effort. *Journal of Systems and Software*, 70(1-2):37–60, 2004.

[Jr.95] Frederick P. Brooks Jr. *The Mythical Man Month: Essays on Software Engineering*. Addison-Wesley, 2 edition, 1995.

[Kem87] Chris F. Kemerer. An empirical validation of software cost estimation models. *Communications of the ACM*, 30(5):416–429, 1987.

[Lar90] Luiz A. Laranjeira. Software size estimation of object-oriented systems. *IEEE Transactions on Software Engineering*, 16(5):510–522, Maio 1990.

[lEG07] Álvaro Eduardo Gomes. Métricas e estimativas de software - o início de um rally de regularidade. *APInfo Website*, 2007.

[Lov86] M. C. Lovell. Tests of the rational expectations hypothesis. *The American Economic Review*, 1986.

[LP92] Albert L. Lederer and Jayesh Prasad. Nine management guidelines for better cost estimating. *Communications of the ACM*, pages 51–59, 1992.

[Mad01] G. S. Maddala. *Introduction to Econometrics*. John Wiley & Sons, Ltd., 2001.

[McC06] Steve McConnell. *Software Estimation – Demystifying the Black Art*. Microsoft Press, 2006.

[Mil83] Harlan D. Mills. *Software Productivity*. Little, Brown, Boston, MA, 1983.

[Par88] R. Park. The central equations of the price software cost model. In *4th COCOMO Users Group Meeting*, Novembro 1988.

[PM92] Lawrence H. Putnam and Ware Myers. Measures for excellence : reliable software on time, within budget, 1992.

[(PM97a] Project Management Institute (PMI). A guide to the project management body of knowledge (pmbok), August 1997.

[PM97b] Lawrence H. Putnam and Ware Myers. Industrial strength software: Effective management using measurement, 1997.

[PM03] Lawrence H. Putnam and Ware Myers. Five core metrics. Dorset House, 2003.

[Pre00] Lutz Prechelt. An empirical comparison of seven programming languages. *IEEE Computer*, 33(10):23–29, 2000.

[Ros82] Kenneth T. Rosen. The impact of proposition 13 on housing prices in northern california: A test of the interjurisdictional capitalization hyphothesis. *Journal of Political Economy*, pages 191–200, February 1982.

[SEG68] H. Sackman, W. J. Erikson, and E. E. Grant. Exploratory experimental studies comparing online and offline programming performance. *Commun. ACM*, 11(1):3–11, 1968.

[Stu05] Richard D. Stutzke. *Estimating Software-Intensive Systems*. Addison-Wesley, Upper Saddle River, NJ, 2005.

[VM89] Jon D. Valett and Frank E. McGarry. A summary of software measurement experiences in the software engineering laboratory. *Journal of Systems and Software*, 9(2):137–148, 1989.

[Wau64] F. V. Waugh. Demand and price analysis: Some examples from agriculture. Technical report, U. S. D. A., 1964.

[Web07] Website. International function point users group (http://www.ifpug.org/), Junho 2007.

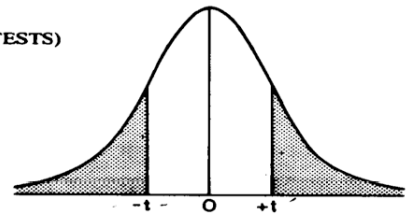
[Web08] Website. Dr. michael thomas flanagan's home page (http://www.ee.ucl.ac.uk/ mflanaga/), Junho 2008.

[WS74] Gerald M. Weinberg and Edward L. Schulman. Goals and performance in computer programming. *IEEE Transactions on Software Engineering*, 16:70–77, 1974.

Apêndices

Apêndice A: Tabela t de student

TABLE A 4
DISTRIBUTION OF t (TWO-TAILED TESTS)

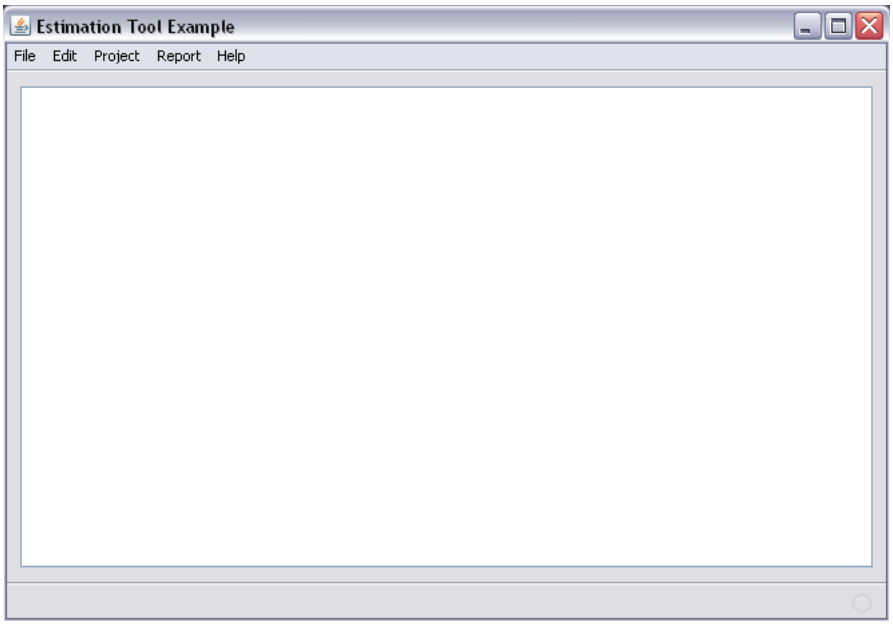


Degrees of Freedom	Probability of a Larger Value, Sign Ignored								
	0.500	0.400	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	1.000	1.376	3.078	6.314	12.706	25.452	63.657		
2	0.816	1.061	1.886	2.920	4.303	6.205	9.925	14.089	31.598
3	.765	0.978	1.638	2.353	3.182	4.176	5.841	7.453	12.941
4	.741	.941	1.533	2.132	2.776	3.495	4.604	5.598	8.610
5	.727	.920	1.476	2.015	2.571	3.163	4.032	4.773	6.859
6	.718	.906	1.440	1.943	2.447	2.969	3.707	4.317	5.959
7	.711	.896	1.415	1.895	2.365	2.841	3.499	4.029	5.405
8	.706	.889	1.397	1.860	2.306	2.752	3.355	3.832	5.041
9	.703	.883	1.383	1.833	2.262	2.685	3.250	3.690	4.781
10	.700	.879	1.372	1.812	2.228	2.634	3.169	3.581	4.587
11	.697	.876	1.363	1.796	2.201	2.593	3.106	3.497	4.437
12	.695	.873	1.356	1.782	2.179	2.560	3.055	3.428	4.318
13	.694	.870	1.350	1.771	2.160	2.533	3.012	3.372	4.221
14	.692	.868	1.345	1.761	2.145	2.510	2.977	3.326	4.140
15	.691	.866	1.341	1.753	2.131	2.490	2.947	3.286	4.073
16	.690	.865	1.337	1.746	2.120	2.473	2.921	3.252	4.015
17	.689	.863	1.333	1.740	2.110	2.458	2.898	3.222	3.965
18	.688	.862	1.330	1.734	2.101	2.445	2.878	3.197	3.922
19	.688	.861	1.328	1.729	2.093	2.433	2.861	3.174	3.883
20	.687	.860	1.325	1.725	2.086	2.423	2.845	3.153	3.850
21	.686	.859	1.323	1.721	2.080	2.414	2.831	3.135	3.819
22	.686	.858	1.321	1.717	2.074	2.406	2.819	3.119	3.792
23	.685	.858	1.319	1.714	2.069	2.398	2.807	3.104	3.767
24	.685	.857	1.318	1.711	2.064	2.391	2.797	3.090	3.745
25	.684	.856	1.316	1.708	2.060	2.385	2.787	3.078	3.725
26	.684	.856	1.315	1.706	2.056	2.379	2.779	3.067	3.707
27	.684	.855	1.314	1.703	2.052	2.373	2.771	3.056	3.690
28	.683	.855	1.313	1.701	2.048	2.368	2.763	3.047	3.674
29	.683	.854	1.311	1.699	2.045	2.364	2.756	3.038	3.659
30	.683	.854	1.310	1.697	2.042	2.360	2.750	3.030	3.646
35	.682	.852	1.306	1.690	2.030	2.342	2.724	2.996	3.591
40	.681	.851	1.303	1.684	2.021	2.329	2.704	2.971	3.551
45	.680	.850	1.301	1.680	2.014	2.319	2.690	2.952	3.520
50	.680	.849	1.299	1.676	2.008	2.310	2.678	2.937	3.496
55	.679	.849	1.297	1.673	2.004	2.304	2.669	2.925	3.476
60	.679	.848	1.296	1.671	2.000	2.299	2.660	2.915	3.460
70	.678	.847	1.294	1.667	1.994	2.290	2.648	2.899	3.435
80	.678	.847	1.293	1.665	1.989	2.284	2.638	2.887	3.416
90	.678	.846	1.291	1.662	1.986	2.279	2.631	2.878	3.402
100	.677	.846	1.290	1.661	1.982	2.276	2.625	2.871	3.390
120	.677	.845	1.289	1.658	1.980	2.270	2.617	2.860	3.373
∞	.6745	.8416	1.2816	1.6448	1.9600	2.2414	2.5758	2.8070	3.2905

Apêndice B: Manual do sistema implementado

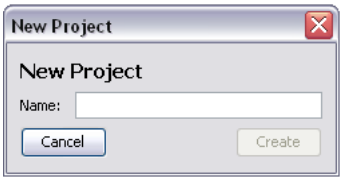
Tela Principal

A tela principal do sistema é composta de duas partes, uma barra superior, por onde todas as funcionalidades do sistema podem ser acessadas, e uma caixa de texto, onde todas as mensagens geradas pelo sistema serão exibidas.



New Project

Acessando o menu File, temos a opção "New Project". Esta opção adiciona um novo projeto, com o nome especificado, ao sistema.



O campo "Name" é obrigatório.

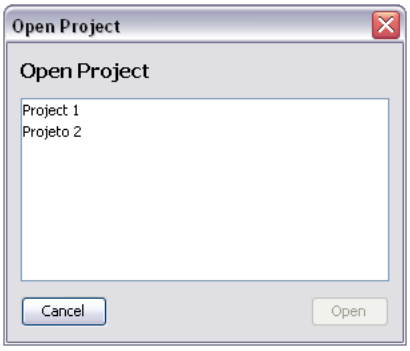
Assim que o projeto é adicionado, o sistema passa a adotá-lo como projeto atual, e todas as alterações de projeto serão feitas sobre ele.

Caso a ação seja efetuada com sucesso, o sistema deverá apresentar a seguinte mensagem:

```
Project 'project_name' successfully created!
```

Open Project

Acessando o menu File, temos a opção "Open Project". Esta opção carrega um projeto já criado no sistema.



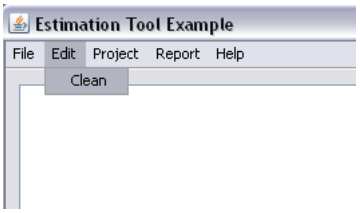
Assim que o projeto é carregado, o sistema passa a adotá-lo como projeto atual, e todas as alterações de projeto serão feitas sobre ele.

Quando efetuada com sucesso, a ação deve retornar a seguinte mensagem:

```
Project 'project_name' successfully opened!
```

Clean

Acessando o menu Edit, temos a opção Clean. Esta opção ativa a função de limpeza da area de texto presente na tela principal.

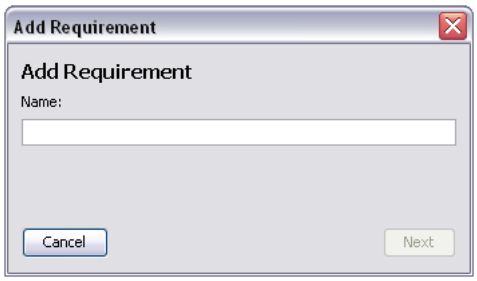


Add Requirement

A opção "Add Requirement" adiciona um novo requisito ao sistema associado ao projeto atual, devido a isto, é necessário que um projeto esteja aberto, caso contrário, o sistema deverá retornar a seguinte mensagem:

```
Cannot Add Requirement to Project!
No Project Opened.
```

A primeira tela apresentada para criação de um novo requisito é a seguinte:



Onde é solicitado o nome do novo requisito. O campo "Name" é obrigatório.

A próxima tela apresentada, solicita quais tarefas serão realizadas no requisito que está sendo criado.

O usuário deve selecionar todas as tarefas que deverão ser estimadas para o novo requisito.

Quando efetuada com sucesso, a ação deve retornar a seguinte mensagem:

```
Requirement 'requirement_name' successfully added!
Task 'Coding and Unit Test' successfully added to requirement 'requirement_name'!
Task 'SIT Execution' successfully added to requirement 'requirement_name'!
```

Start Phase

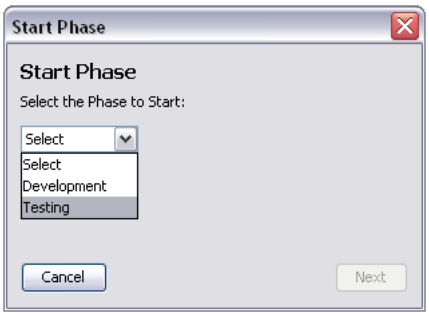
A opção "Start Phase" inicia uma nova fase no projeto atual, solicitando todos os atributos para geração das estimativas. É necessário que um projeto esteja aberto, caso contrário, o sistema deverá retornar a seguinte mensagem:

```
Cannot Start Phase!
No Project Opened.
```

Também é necessário que o projeto possua fases para serem iniciadas, caso contrário, o sistema retornará a seguinte mensagem:

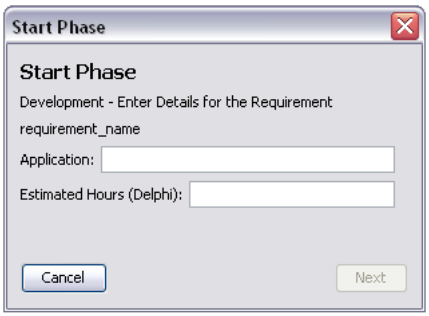
```
Cannot Start Phase!
Project with No Phases to Start.
```

A primeira tela apresentada ao usuário solicita qual fase deverá ser iniciada:

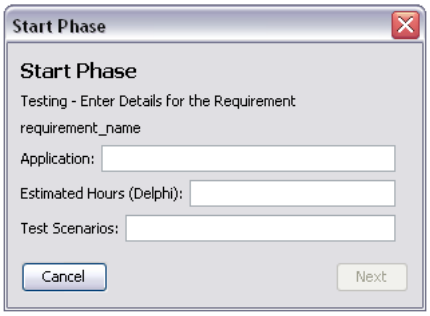


O usuário deverá selecionar a fase que deseja iniciar pressionar o botão “Next”.

As próximas telas apresentadas solicitarão os atributos de cada requisito que está sendo estimado nesta fase. Caso a fase seja “Development”, o sistema apresentará a seguinte tela:



Caso a fase seja “Testing”, o sistema deverá apresentar a seguinte tela:



Assim que os atributos de todos requisitos forem informados, o sistema apresentará a seguinte tela:

Ao pressionar o botão “Start” o sistema irá marcar a fase como iniciada e gerar as estimativas para cada requisito.

Quando efetuada com sucesso, a ação deve retornar a seguinte mensagem:

```
'phase_name' Phase successfully started!
'task_name' Estimates:
Requirement                    Estimate (Hours)
requirement_name               hours
```

Finish Phase

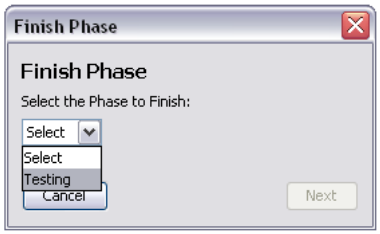
A opção “Finish Phase” finaliza uma fase já iniciada no projeto atual, solicitando o esforço real de cada requisito para criação de novas equações de regressão. É necessário que um projeto esteja aberto, caso contrário, o sistema deverá retornar a seguinte mensagem:

```
Cannot Finish Phase!
No Project Opened.
```

Também é necessário que o projeto possua fases para serem finalizadas, caso contrário, o sistema retornará a seguinte mensagem:

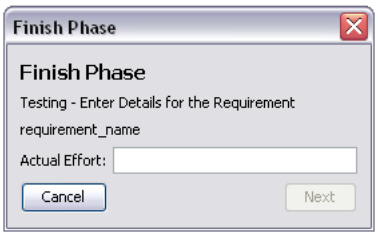
```
Cannot Finish Phase!
Project with No Phases to Finish.
```

A primeira tela apresentada ao usuário solicita qual fase deverá ser finalizada:



O usuário deverá selecionar a fase que deseja finalizar pressionar o botão “Next”.

As próximas telas apresentadas solicitarão o esforço real de cada requisito que estimado nesta fase.



Assim que o esforço real de todos requisitos forem informados, o sistema apresentará a seguinte tela:

Ao pressionar o botão "Finish" o sistema irá marcar a fase como finalizada e retornará a seguinte mensagem:

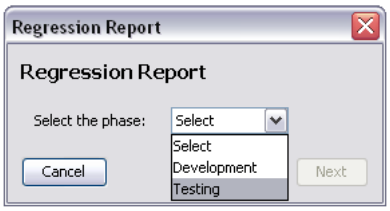
```
'phase_name' Phase Successfully Finished!
```

Regression Report

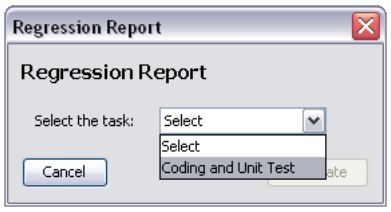
A opção "Regression Report" deverá gerar um relatório de regressão sobre um modelo selecionado. O relatório deverá conter as seguintes informações:

- Melhores Estimativas dos Coeficientes de Regressão
- Coeficientes de Correlação Linear
- Valores Utilizados na Regressão
- Valores Calculados pela Regressão
- Resíduos
- Média de Resíduos
- Graus de Liberdade
- Número de Parâmetros de Entrada
- Número de Parâmetros Estimados

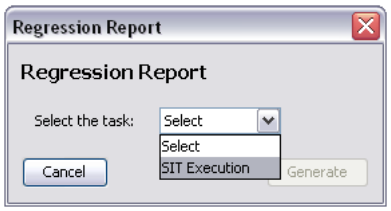
A primeira tela apresentada requisita ao usuário a qual fase pertence a modelo que deve ser analisado:



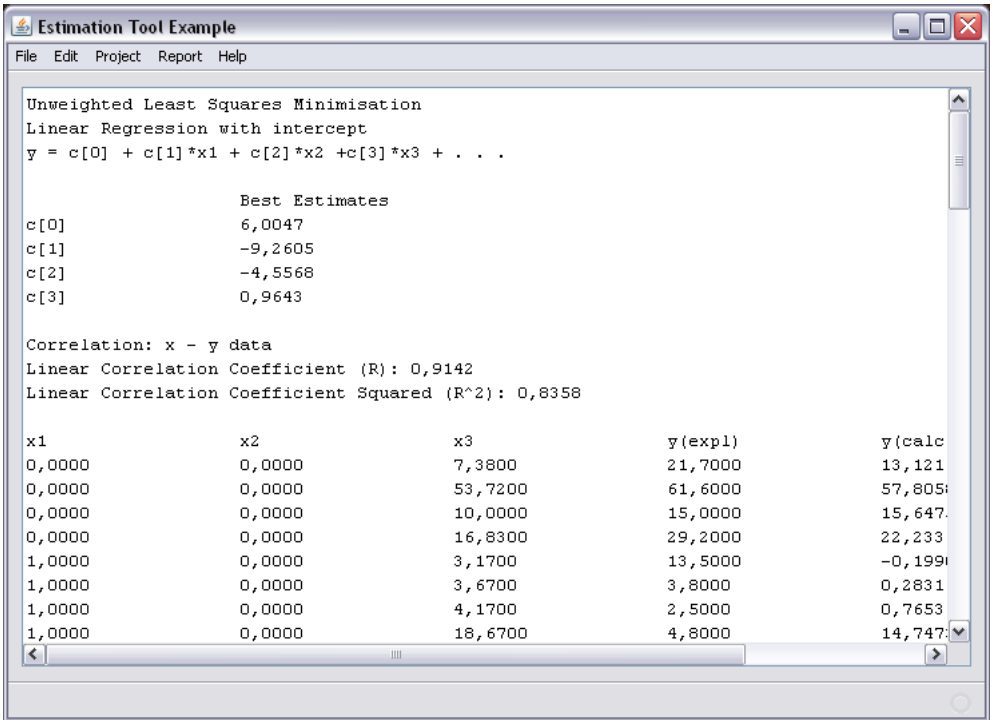
A próxima tela, requisita ao usuário qual tarefa representa o modelo que deve ser analisado. Caso a fase selecionada anteriormente seja “Development” a tela apresentada será esta:



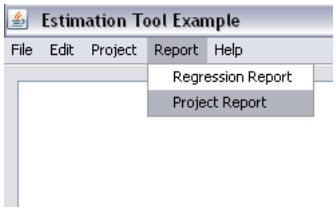
Caso seja “Testing”, a tela apresentada será esta:



Ao pressionar o botão “Generate” o sistema deverá exibir o relatório na caixa de texto presente na tela principal.



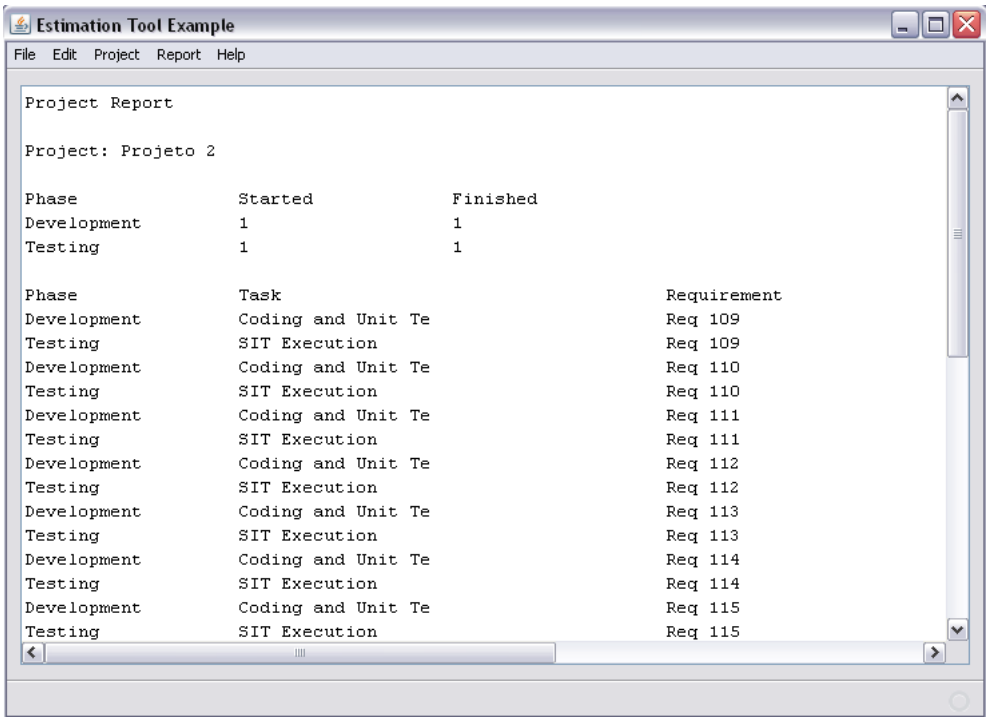
Project Report



Esta opção deverá gerar um relatório sobre o projeto atual, contendo as seguintes informações:

- Nome do Projeto
- Fases do Projeto
- Fases Iniciadas
- Fases Finalizadas
- Tarefas do Projeto
- Esforço Estimado para cada Tarefa
- Esforço Real de cada Tarefa

- Erro entre Esforço Estimado e Real



Esta opção trabalha sobre o projeto atual do sistema, portanto um projeto deve estar aberto para que o relatório seja gerado, caso contrário, o sistema exibirá a seguinte mensagem:

```
Cannot generate Project Report!
No Project Opened.
```