Regressao Simples

Parte I: Introdução

Curso

- A aplicação da análise de regressão requer conhecimento teórico e experiência com análise de dados.
- Este curso procura
  - combinar a teoria estatística com a prática, dando mais enfâse na aplicação.
  - cobrir não apenas tópicos convencionais na regressão, mas por exemplo, modelos de regressão com variáveis preditoras qualitativas, regressão não linear.

Termo Regressão

- O termo regressão foi empregado pela primeira vez por Francis Galton (1822-1911) num estudo da relação entre as alturas dos pais e filhos.
- Os modelos de regressão são largamente utilizados em todas as áreas do conhecimento, tais como: computação, administração, engenharias, biologia, agronomia, saúde, sociologia, etc.
- Os modelos de regressão usam variáveis contínuas ou discretas como sendo as de interesse ou objetivo (*target*). Desejamos aproximar uma função de variáveis de entrada *(inputs)* aos dados.
- O principal objetivo dos modelos de regressão é modelar o relacionamento entre diversas variáveis preditoras e uma variável resposta. Este relacionamento pode ser por uma equação linear ou uma função não linear.

Uso de Regressão em Mineração de Dados

- Extrair modelos de interesse em grandes bancos de dados. Os modelos
- São chamados de modelos preditivos, isto é, são usados para predizer um valor numérico contínuo que está associado a um registro do banco de dados. Exemplo: o tempo até o abandono de um determinado cliente (Análise de sobrevivência) ou a probabilidade do cliente saldar um empréstimo (regressão logística).
- É uma área (regressão e *data mining*) recente, onde pode ser desenvolvida pesquisa no sentido de melhorar a performance dos métodos tradicionais de regressão em grandes bases de dados. Pode-se aplicar as metodologias já desenvolvidas.

Relações

- Todos os dias, a *mídia* se encarrega de informar resultados de análises e pesquisas do tipo:
  - O valor da empresa depende do lucro futuro, a taxa de juro depende da inflação.
  - O salário depende da escolaridade do trabalhador etc.

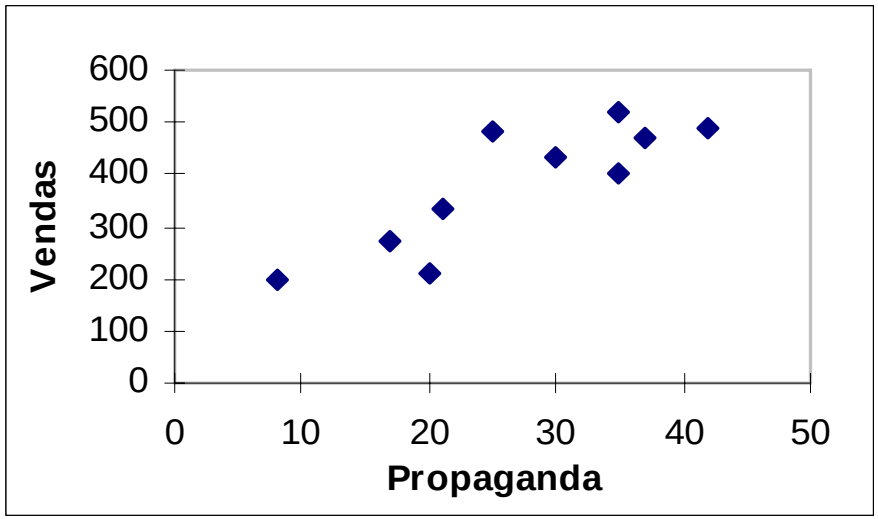
Exemplo

- O objetivo do diretor de vendas de uma rede de varejo é analisar a relação entre o investimento realizado em propaganda e as vendas das lojas da rede, para realizar projeções de vendas de futuros investimentos em propaganda.
- A tabela seguinte registra uma amostra representativa extraída dos registros históricos das lojas de tamanho equivalente, com os valores de Propaganda e Vendas em milhões.
- Analisar a possibilidade de definir um modelo que represente a relação entre as duas variáveis ou amostras.

Propaganda	30	21	35	42	37	20	8	17	35	25
Vendas	430	335	520	490	470	210	195	270	400	480

Solução

- Para analisar a relação entre as duas variáveis foi construído o gráfico de dispersão das vendas anuais em função do investimento anual em propaganda. Nesse gráfico pode-se ver que, nos últimos dez anos, o aumento de investimento em propaganda gerou aumento das vendas, e vice-versa.



Solução

- O objetivo deste Exemplo é ajustar uma reta a partir dos valores das amostras retiradas da população, considerando que o investimento em propaganda é a variável independente *x,* e as vendas anuais, a variável dependente *y*.
- Uma primeira forma de fazer isso é ajustar manualmente essa reta tentando equilibrar os pontos acima e abaixo dessa reta, como foi feito no gráfico deste Exemplo.
- Como esse procedimento permite o ajuste de diversas retas, é necessário estabelecer um objetivo de eficiência de ajuste possível de medir, como é mostrado a seguir.

Modelo do Ajuste de uma Reta

- O ajuste de uma reta é um modelo linear que relaciona a variável dependente *y* e a variável independente *x* por meio da equação de uma reta do tipo:

$$y = \beta_0 + \beta_1 x$$

- É importante observar que os pontos não caem exatamente na reta e a equação deve ser modificada. A diferença é chamada de erro é uma variável aleatória

$$e = y - (\beta_0 + \beta_1 x)$$

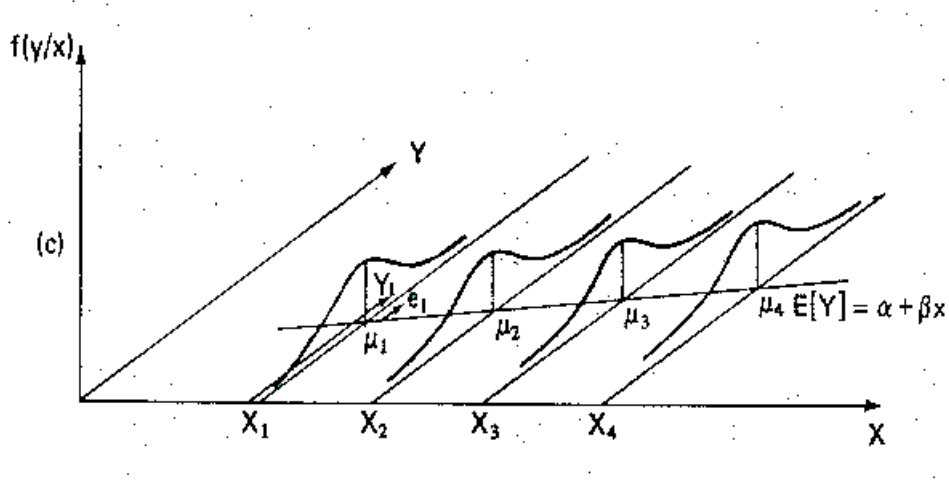
- Então o modelo mais pláusível é: $y = \beta_0 + \beta_1 x + e$
- Se fixarmos o valor de *x* e observar um valor de *y* e supormos que o erro *e* é um componente aleatório com média zero e variancia constante σ^2, média da variável resposta dado x é

$$E(y/x) = E(\beta_0 + \beta_1 x + e) = \beta_0 + \beta_1 x$$

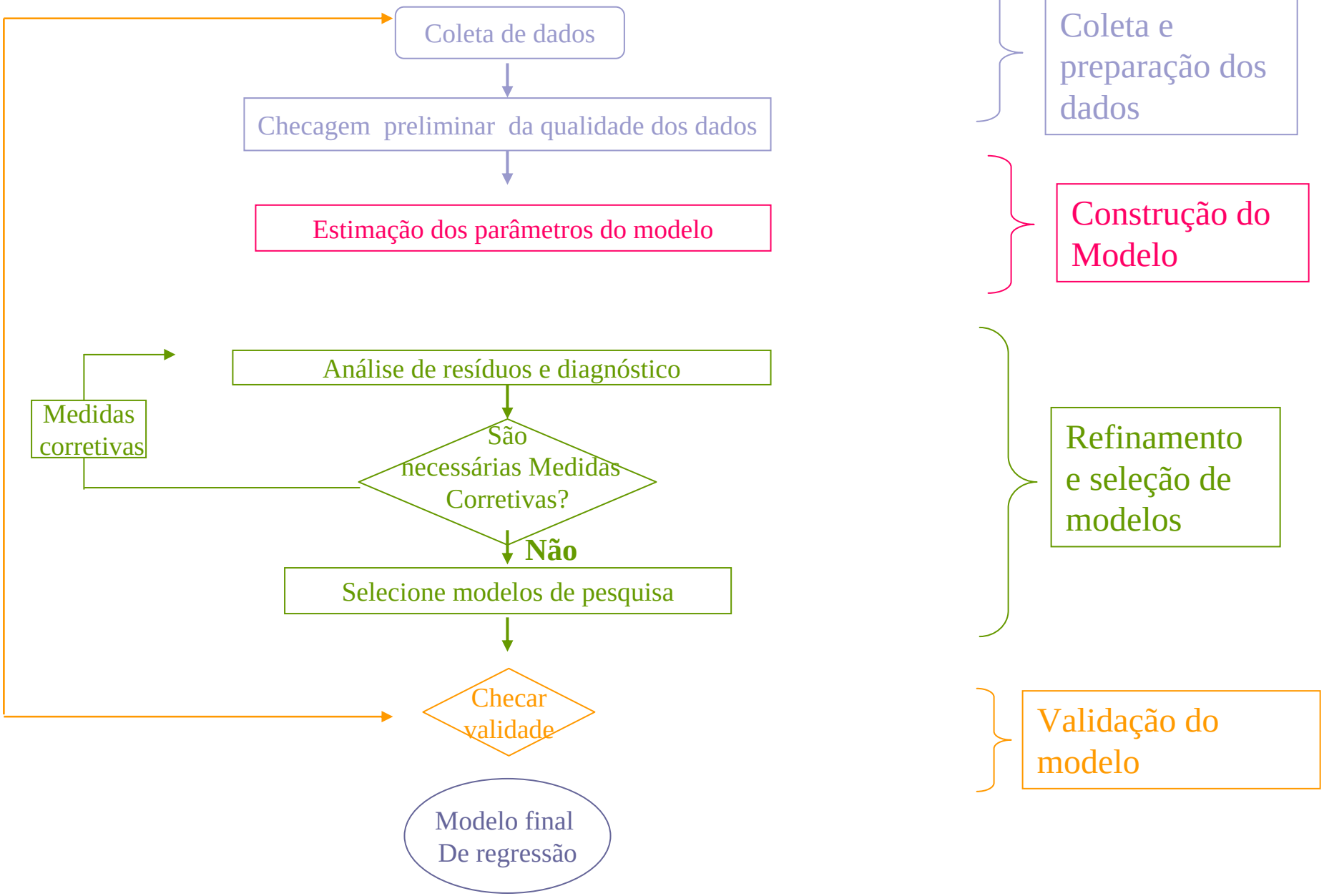
- e a variância de *y* dado *x* é:

$$V(y/x) = V(\beta_0 + \beta_1 x + e) = \sigma^2$$

Exemplo



Existe uma distribuição de y a cada valor de x e a variância dessa distribuição é constante


Coleta de dados
Checagem preliminar da qualidade dos dados
Estimação dos parâmetros do modelo
Análise de resíduos e diagnóstico
Medidas corretivas
São necessárias Medidas Corretivas?
Não
Selecione modelos de pesquisa
Checar validade
Modelo final De regressão
Coleta e preparação dos dados
Construção do Modelo
Refinamento e seleção de modelos
Validação do modelo

Coleta e Preparação dos dados

Coletados os dados → Organizar, resumir, explorar,

↓

Verificar erros grosseiros, outliers

↓

Os erros devem ser corrigidos antes de iniciar a construção do modelo (crítico em grandes bases de dados)

←

Sempre que possível o pesquisador deve estar presente na coleta dos dados

Métodos de Coleta

- Dados históricos:
  - São dados que nunca foram coletados
  - Informações convenientes para coletar
  - Sofre erros de transcrição
- Estudos observacionais
  - Geralmente *X* e *Y* são variáveis aleatórias.
  - Assegura acurácia e confiabilidade
  - Minimiza outlier
  - Colinariedade
- Experimental
  - Caracterizam-se por apresentarem manipulação de intervenções diretas sobre os indivíduos em estudo. Exemplo típico ensaio clínico.
  - Geralmente *X* (doses, tempo) é determinado pelo pesquisador $\rightarrow$ *X* é fixo. *Y* está sujeito à variações físicas, biológicas, de medidas, etc. $\rightarrow$ *Y* é uma variável aleatória.
  - Elimina colinariedade

Considerações

- *Dois modelos:*
  - *X fixo:* ajusta-se um modelo para a média da v. aleatória *Y* como uma função de *X* fixo (linha reta). Estima-se os parâmetros do modelo para caracterizar o relacionamento.
  - *X aleatório:* caracteriza-se o relacionamento (linear) entre *X e Y* através da correlação entre elas e estima-se o parâmetro de correlação.
- *Sutileza:* em situações onde *X* é uma variável aleatória, muitos investigadores desejam ajustar um modelo de regressão tratando *X* como fixo. Isto porque, embora o coef. de correlação descreve o grau de associação entre *X e Y*, ele não caracteriza o relacionamento através de um modelo de regressão.
  - Se *X e Y* são variáveis aleatórias, e nós ajustarmos um modelo de regressão para caracterizar o relacionamento, tecnicamente, todas as análises posteriores são consideradas como sendo **condicionais** aos valores de *X* presentes no estudo. Isto significa que nós consideramos *X* fixo, embora ele não seja. Entretanto, é válido fazer-se previsões. **Dado (condicional)** que se observa um particular valor de altura de planta, ele quer obter o melhor valor para produção. O pesquisador não está dizendo que ele pode controlar as alturas e, assim, influenciar as produções.

Cronstrução do modelo

Dados fidedignos

↓

Pensar na construção do modelo → Diagnósticos:

1) a forma funcional de como as variáveis preditoras devem entrar no modelo de regressão;

2) interações importantes que devem ser incluídas no modelo.

Diagrama de dispersão, ajuste de funções de regressão para verificar relacionamentos, interações, necessidade de transformações. Usar a experiência do investigador.

Considerações

- Um modelo com poucas variáveis é mais fácil de trabalhar e entender
- A presença de variáveis correlacionadas implica num acréscimo da variância amostral dos coeficientes de regressão, diminuindo a capacidade preditiva e, piorando, também o poder descritivo
- A capacidade preditiva do modelo diminui quando variáveis explanatórias não relacionadas com a variável resposta são mantidas no modelo, dado que as outras variáveis explanatórias estão no modelo.
- A eliminação de variáveis preditoras imprescindíveis prejudica o modelo.
- O modelo que contiver mais variáveis do que o necessário a variância das estimativas dos parâmetros será grande em comparação com modelos mais simples.

Refinamento do Modelo

Gráfico de resíduos versus interações e/ou termos quadráticos, cúbicos, etc. ainda não incluídos no modelo; é útil para identificar termos que podem melhorar o ajuste do modelo.

Aplica regressão stepwise

(processo automático de seleção)

→ Verificar o número de v. regressoras retidas no modelo.

↓ Encontrar outros possíveis modelos

Validação do Modelo

Refere-se a qualidade (estabilidade e razoabilidade) dos coeficientes de regressão; ao bom poder descritivo (plausível, útil) da parte funcional do modelo.

1 - Coleta de novos dados (dados independentes) para checar o modelo e seu poder preditivo;

2 - Comparar os resultados com valores teóricos esperados, resultados empíricos anteriores ou resultados simulados;

3 - Usar uma parte da amostra para checar o modelo e o poder preditivo do mesmo.

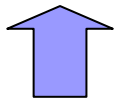
Modelo: sem especificação de distribuição de normalidade

- Considere o modelo com uma variável preditora e que a função de regressão é linear. O modelo é dado por:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

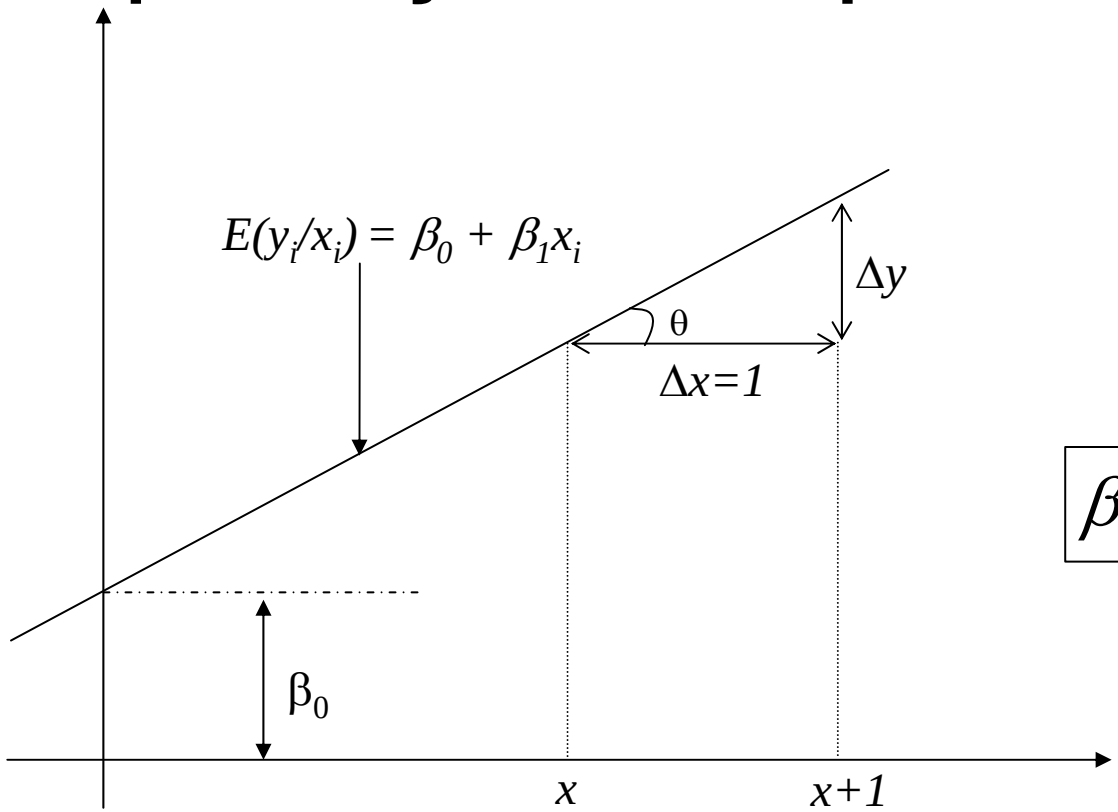
$$i = 1,2,...,n$$

- *Yi* é o i-ésimo valor da variável resposta;
- $\beta 0$ e $\beta 1$ são os parâmetros (coeficientes de regressão);
- *Xi* é o i-ésimo valor da variável preditora (é uma constante conhecida, fixo).
- εi é o termo do erro aleatório com $E(\varepsilon i)=0$ e $\sigma 2(\varepsilon i)= \sigma 2$;
- εi e εj não são correlacionados $\Rightarrow \sigma(\varepsilon i, \varepsilon j)=0$ para todo $i,j; i \neq j;$ (covariância é nula).
- $i=1,2,...,n.$



Covariância (o resultado em qualquer experimento não tem efeito no termo do erro de qualquer outro experimento)

Interpretação dos parâmetros



$$\beta_1 = \frac{\Delta y}{\Delta x}$$

β_0 (intercepto); quando a região experimental inclui *X=0,* β_0 é o valor da média da distribuição de Y em X=0, não tem significado prático como um termo separado (isolado) no modelo;

β_1 (inclinação) expressa a *taxa de mudança* em *Y,* isto é, é a mudança em *Y* quando ocorre a mudança de uma unidade em *X.* Ele indica a mudança na média da distribuição de probabilidade de *Y* por unidade de acréscimo em *X.*

Método dos Mínimos Quadrados

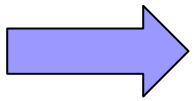
- De acordo com o método de mínimos quadrados, os estimadores de β_0 e β_1 são os valores $\hat{\beta}_0$ e $\hat{\beta}_1$, respectivamente, que minimizam o critério S para a amostra $(X_1,Y_1),..,(X_n,Y_n)$.

$$S(\beta_0,\beta_1)=\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1 X_i\right)^2$$

- Derivando e igualando a zero

$$\frac{\partial S}{\partial \beta_0}=-2\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1 X_i\right)=0$$

$$\frac{\partial S}{\partial \beta_1}=-2\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1 X_i\right)x_i=0$$



$$n\hat{\beta}_0+\hat{\beta}_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i$$

$$\hat{\beta}_0\sum_{i=1}^{n}x_i+\hat{\beta}_1\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}y_i x_i$$

Método dos Mínimos Quadrados

A solução para as duas equações normais são:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\left(\sum_{i=1}^{n} x_i^2\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)}{n}} \quad \Rightarrow \quad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

$$S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{XY} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

Modelo ajustado $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Resíduo $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x)$

Propriedades dos estimadores de mínimos quadrados

- São combinações lineares das observações.
- São estimadores não enviesados.
- Teorema de Gauss-Markov: assumindo erro com média zero e variância constante para o modelo de regressão, os estimadores são não enviesados e de variância mínima comparado com os outros estimadores não tendenciosos que são combinações lineares das observações.
- A soma dos resíduos do modelo que contém o intercepto é sempre zero.
- A soma dos valores observados de Y são iguais a soma dos valores ajustados de Y.
- A linha de regressão sempre passa através do centroide $(\bar{x}, \bar{y})$ e

$$\sum_{i=1}^{n} x_i e_i = 0 \qquad \sum_{i=1}^{n} \hat{y}_i e_i = 0$$

Estimativa de σ^2

- É usada para construir intervalos de confiança pertinente ao modelo de regressão. Idealmente isso é feito independentemente da adequacidade do modelo ajustado. Isto é quando existem diversas observações de y para pelo menos um valor de x. Quando esta abordagem não pode ser obtida então estima-se pela soma de quadrados dos resíduos

$$SS_{\text{Res}} = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Podemos reescrever essa fórmula por $SS_{\text{Res}} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{XY}$
- Mas $\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SST$
- Assim $SS_{\text{Res}} = SS_T - \hat{\beta}_1 S_{XY}$
- Tem-se que $E(SS_{\text{Res}}) = (n-2)\sigma^2 \Rightarrow \hat{\sigma}^2 = \frac{SS_{\text{Res}}}{n-2} = MS_{\text{Res}}$

Forma alternativa do modelo

- Suponha que o variável regressora é redefinida $x_i = x_i - \bar{x}$
- A fim de manter os mesmos valores ajustados, o modelo de regressão é

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_1\bar{x} + \varepsilon_i = (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + \varepsilon_i$$

$$y_i = \beta_0' + \beta_1(x_i - \bar{x}) + \varepsilon_i$$

- A relação entre o interceptos original e o novo é $\beta_0' = \beta_0 + \beta_1\bar{x}$
- O estimador da inclinação não é afetado.
- Além disso os estimadores são não correlacionados. Isso torna algumas aplicações do modelo mais fáceis tal como encontrar intervalos de confiança da média de Y.
- Finalmente, o modelo ajustado é

$$\hat{y} = \bar{y} + \beta_1(x - \bar{x})$$

Testando Hipótese sobre os coeficientes (parâmetros)

- É necessário supor que erros tem distribuição normal, são independentes e tem média zero e variância constante

$$e_i \rightarrow NID(0, \sigma^2)$$

- Assim os valores observados y_i são normalmente e independente distribuídos com média e variância constante

$$y_i \rightarrow NID(\beta_0 + \beta_1 x_i, \sigma^2)$$

- Assim temos

$$\hat{\beta}_1 \rightarrow ND(\beta_1, \sigma^2 / S_{XX})$$

- Hipóteses

$$H_0 : \beta_1 = \beta_{10}$$

$$H_a : \beta_1 \neq \beta_{10}$$

- Usando os resultados acima temos

$$Z = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2 / S_{XX}}} \rightarrow N(0,1)$$

- Se a variância não é conhecida

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{\text{Res}} / S_{XX}}} \rightarrow t_{n-2}$$

Erro padrão da estimativa

- Rejeita H_0 se $|t_0| > t_{\alpha/2, n-2}$

Testando Hipótese sobre os coeficientes (parâmetros)

- Hipóteses

$$H_0 : \beta_0 = \beta_{00}$$
$$H_a : \beta_0 \neq \beta_{00}$$

- Se a variância não é conhecida

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{\text{Res}}(1/n + \bar{x}^2 / S_{XX})}} \rightarrow t_{n-2}$$

Erro padrão do intercepto

- Rejeita H_0 se

$$| t_0 | > t_{\alpha/2, n-2}$$