IAA005 - ESTATÍSTICA APLICADA I
Parte 3

Prof. Arno P. Schmitz

UFPR – Universidade Federal do Paraná

Análise de Regressão Linear

A análise de regressão linear baseia-se em um modelo matemático teórico, como por exemplo a função consumo das famílias:

$$Y = \beta_1 + \beta_2 X$$

Em que:

Y = Despesas de consumo;

X = Renda disponível;

β_1 e β_2 = Parâmetros a serem estimados (intercepto e coeficiente angular, respectivamente).

➔β_2 mede a propensão marginal a consumir da renda disponível;

➔β_1 é interpretado como o consumo autônomo que independe da renda;

Análise de Regressão Linear

Para qualquer modelo matemático teórico exemplificado abaixo:

$$Y = \beta_1 + \beta_2 X$$

- A variável que aparece do lado esquerdo da igualdade é chamada de **“variável dependente”**;

- As variáveis do lado direito são chamadas de **“variáveis independentes”** ou **“variáveis explanatórias”**, bem como alguns outros nomes;

- Os β_s são parâmetros a serem estimados.

Análise de Regressão Linear

O modelo matemático exige uma relação exata ou determinística entre as variáveis. Mas as relações entre variáveis econômicas e sociais são em geral, inexatas.

- Portanto, se coletarmos dados sobre despesas de consumo e renda disponível (a renda depois de descontados os impostos) de uma amostragem de, digamos, 500 famílias e traçarmos um gráfico em que o eixo vertical representa as despesas de consumo e, o eixo horizontal a renda disponível, não devemos esperar que as 500 observações se situem exatamente sobre a reta dada pela Equação.
- Isto porque, além da renda, outras variáveis afetam o consumo tais como: tamanho da família, idade dos componentes da família, religião, localização geográfica, etc.

Análise de Regressão Linear

- Para considerar as relações inexatas entre as variáveis deve-se ter em mente um modelo estatístico ou econométrico, tal como:

$$Y = \beta_1 + \beta_2 X + u$$

- Este modelo econométrico acima é o caso de regressão simples, pois existe uma variável dependente e apenas uma variável explicativa. Mas é possível ter uma modelo de regressão múltipla, que considera duas ou mais variáveis explicativas:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + u$$

- Para resolver o modelo deve-se dispor dos valores das variáveis em um período ou espaço determinado.
- Em outras palavras deve-se ter em mãos uma matriz com os valores das variáveis.

Análise de Regressão Linear

Ano	DCP(*Y*)	PIB(*X*)
1960	1597,4	2501,8
1961	1630,3	2560,0
1962	1711,1	2715,2
1963	1781,6	2834,0
1964	1888,4	2998,6
1965	2007,7	3191,1
1966	2121,8	3399,1
1967	2185,0	3484,6
1968	2310,5	3652,7
1969	2396,4	3765,4
1970	2451,9	3771,9
1971	2545,5	3898,6
1972	2701,3	4105,0
1973	2833,8	4341,5
1974	2812,3	4319,6
1975	2876,9	4311,2
1976	3035,5	4540,9
1977	3164,1	4750,5
1978	3303,1	5015,0
1979	3383,4	5173,4
1980	3374,1	5161,7
1981	3422,2	5291,7
1982	3470,3	5189,3
1983	3668,6	5423,8
1984	3863,3	5813,6
1985	4064,0	6053,7
1986	4228,9	6263,6
1987	4369,8	6475,1
1988	4546,9	6742,7
1989	4675,0	6981,4
1990	4770,3	7112,5
1991	4778,4	7100,5
1992	4934,8	7336,6
1993	5099,8	7532,7
1994	5290,7	7835,5
1995	5433,5	8031,7
1996	5619,4	8328,9
1997	5831,8	8703,5
1998	6125,8	9066,9
1999	6438,6	9470,3
2000	6739,4	9817,0
2001	6910,4	9890,7
2002	7099,3	10048,8
2003	7295,3	10301,0
2004	7577,1	10703,5
2005	7841,2	11048,6

Análise de Regressão Linear

- A matriz de dados apresentada tem seus dados que podem ser classificados como dados temporais, ou uma série temporal. Isto porque apresentada dados de 1969 a 2005, para dados de consumo - DCP(Y) - e renda bruta da sociedade – PIB(X).

Portanto, o modelo estatístico deve ser corretamente apresentado como:

$$Y_t = \beta_1 + \beta_2 X_t$$

O subscrito "t" apresenta o modelo como sendo um modelo de série temporal.

Alternativamente, um modelo estatístico pode ser apresentado como do tipo cross-section, ou seja, cujos dados são apresentados todos em um mesmo período de tempo e desfrutam da mesma localização geográfica. Este modelo pode ser expresso por:

$$Y_i = \beta_1 + \beta_2 X_i$$

Ou seja, o subscrito "i" apresenta o modelo como sendo do tipo cross-section (corte temporal). Exemplos destes dados podem uma função que deseja saber a produtividade média das máquinas em uma linha de produção e se utiliza dos seguintes dados: produtividade da máquina e gastos com manutenção.

Análise de Regressão Linear

- Um outro tipo de modelo estatístico é aquele no qual os dados estão dispostos segundo sua disposição geográfica, ou seja os dados carregam consigo a informação da sua localização (latitude e longitude).

$$Y_{ui,vi} = \beta_1 + \beta_2 X_{ui,vi}$$

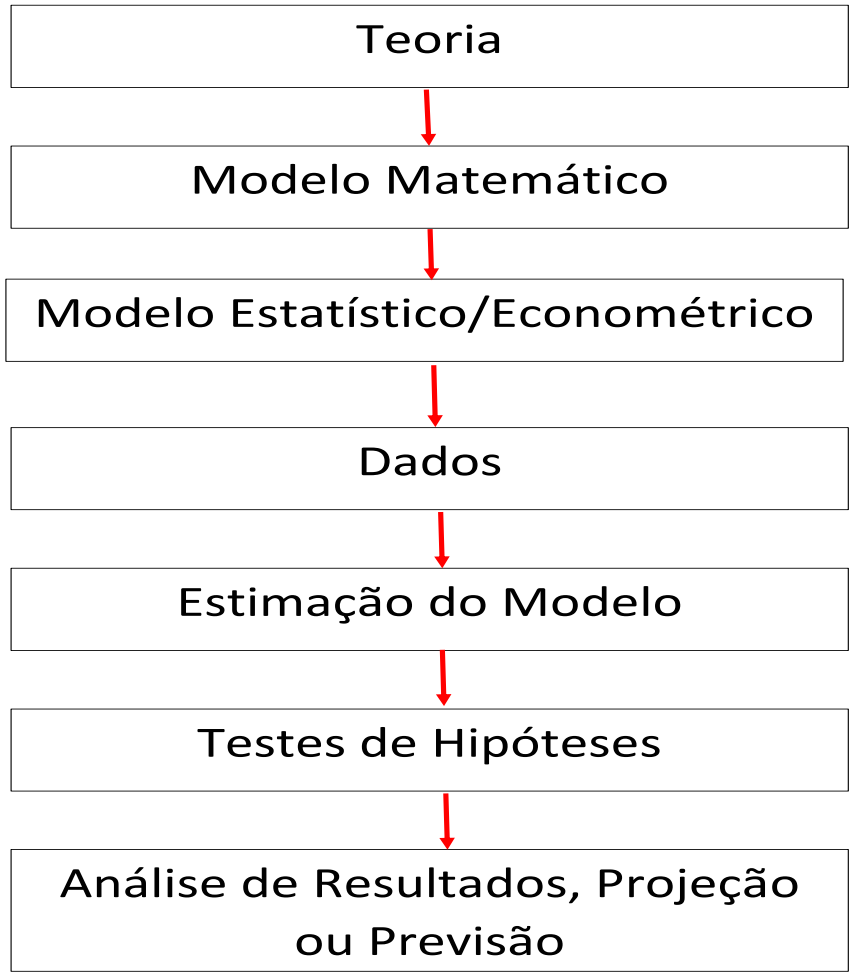
Neste caso, os subscritos ui e vi representam respectivamente a latitude e longitude da observação i, que vai de 1 a k. Sendo assim poderemos ter k observações geograficamente distribuídas.

Deve-se notar que, para o modelo acima, os parâmetros são genéricos para toda a distribuição espacial. Mas, também é possível ter modelos com parâmetros espacialmente distribuídos tal como no seguinte modelo:

$$Y_{ui,vi} = \beta_{1;ui,vi} + \beta_{2;ui,vi} X_{ui,vi}$$

Escolha da Estrutura Inicial do Modelo

- Parte-se de um problema a ser tratado, para tanto deve buscar primeiramente uma teoria que forneça uma base de conhecimento para elaborar o modelo e fazer as análises necessárias.



Escolha da Estrutura Inicial do Modelo

- Contudo, se não houver uma teoria plausível para o problema a ser tratado, pode-se utilizar de experimentação, intuição científica e evidências coletadas em bases de dados.
- Neste caso, elimina-se a necessidade de uma teoria subjacente ao problema de pesquisa.
- Os modelos econométricos tratam da dependência de uma variável em relação a outras, mas não existe necessariamente causação entre as variáveis explicativas e a variável dependente.
- Para ver se uma variável causa impacto em outra ou outras, deve-se utilizar metodologia específica, tal como os testes de causalidade de Granger e outros testes disponíveis na atualidade.

Regressão *versus* Correlação

- A análise de correlação tem como objetivo medir a “força” ou o “grau” de associação linear entre duas ou mais variáveis está estritamente relacionada à análise de regressão, mas conceitualmente é muito diferente.

- O coeficiente de correlação mede a força de associação (linear) – tal como na matriz de correlação.

- Na análise de regressão, busca-se estimar ou prever o valor médio de uma variável com base nos valores fixos de outras variáveis.

- A regressão e a correlação têm algumas diferenças fundamentais. Na análise de regressão, existe uma assimetria na maneira como as variáveis dependente e explanatórias são tratadas. A variável dependente têm uma distribuição de probabilidade. Já para as variáveis explanatórias, considera-se que essas variáveis tem valores fixos em amostras repetidas.

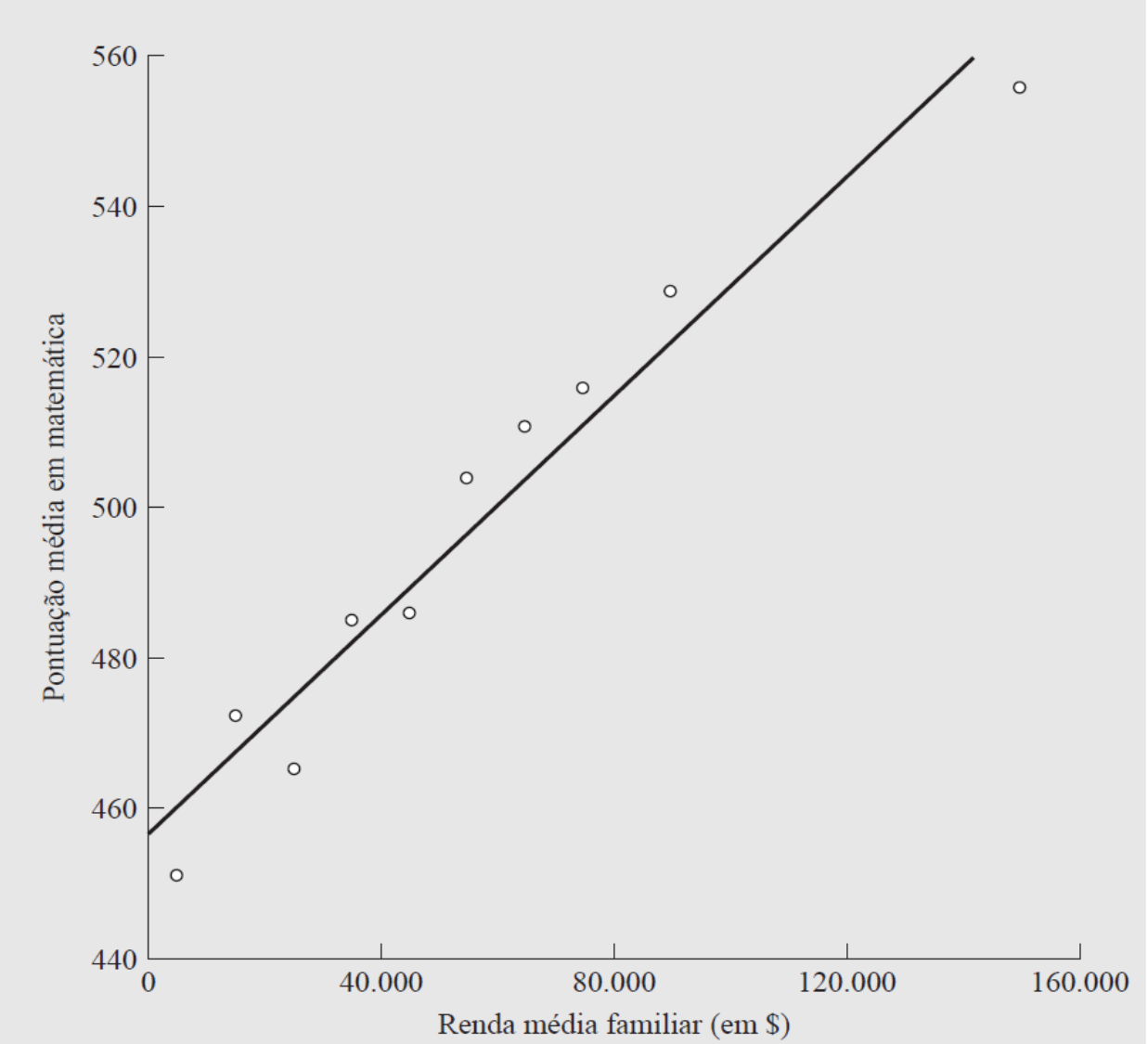
- Na análise de correlação trata-se as variáveis simetricamente, não existe distinção entre variáveis explicativas e dependentes.

Significado do Termo de Erro estocástico (u)

1. Caráter vago da teoria;
2. Indisponibilidade de dados;
3. Variáveis essenciais x Variáveis secundárias;
4. Caráter aleatório do comportamento humano;
5. Varáveis proxy pouco adequadas;
6. Princípio da parcimônia;
7. Forma Funcional errada.

Origem do Termo de Erro estocástico (u)

$Y_i = \beta_1 + \beta_2 X_i + u_i$; Y = Pontuação média matemática , X = Renda média familiar



O Método dos Mínimos Quadrados Ordinários - MQO

Função de Regressão Populacional (FRP):

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

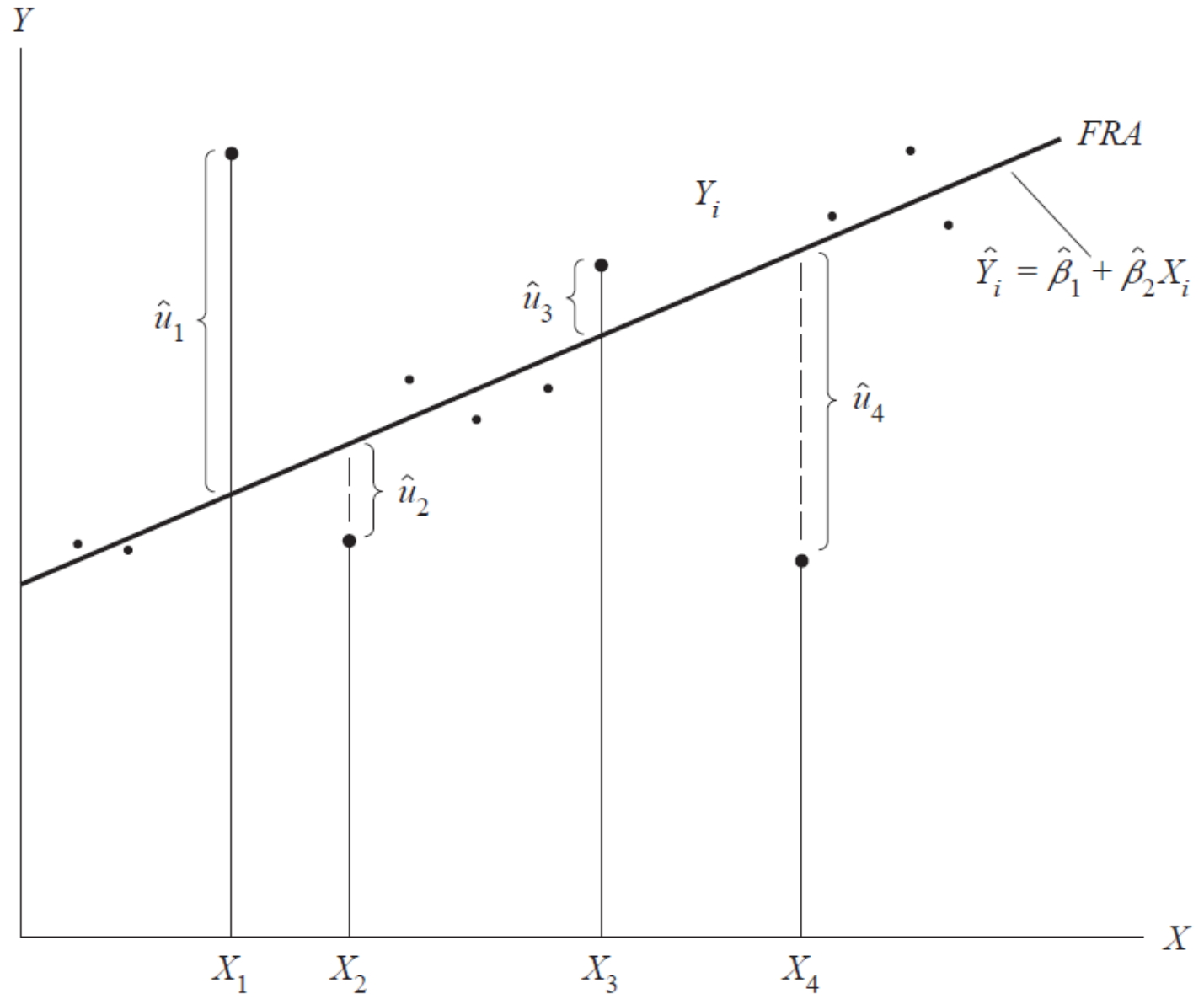
A FRP não pode ser obtida diretamente, portanto deve ser estimada via uma Função de Regressão Amostral (FRA):

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + u_i$$

$$Y_i = \widehat{Y}_i + u_i$$

- As variáveis e parâmetros com “^”, significa que são “estimados”.

O Método dos Mínimos Quadrados Ordinários



O Método dos Mínimos Quadrados Ordinários

Propriedades da Reta de Regressão

1. Passa pelas médias de Y e X;
2. $\bar{\hat{Y}} = \bar{Y}$;
3. A soma dos valores dos resíduos $\hat{u}_i$ é igual a zero $(\sum \hat{u}_i = 0)$;
4. Os resíduos $\hat{u}_i$ não são correlacionados os valores de Y_i;
5. Os resíduos $\hat{u}_i$ não são correlacionados os valores de X_i;

O Método dos Mínimos Quadrados Ordinários

Estimativa do Parâmetro $\hat{\beta}_2$:

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

O Método dos Mínimos Quadrados Ordinários

Estimativa do Parâmetro $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

O Método dos Mínimos Quadrados Ordinários

Hipóteses do Método

1. O modelo de regressão é linear – linearidade nos parâmetros;
2. Os valores de X são fixos em amostras repetidas ou são independentes do termo de erro: valores assumidos pelo regressor X podem ser fixos em amostras repetidas (caso do regressor fixo) ou seus valores podem mudar de acordo com a variável dependente Y (no caso do regressor estocástico). No segundo caso, supõe-se que as variáveis X e o termo de erro são independentes, isto é, $cov(X_i, u_i) = 0$.
3. O valor médio do termo de erro é zero;
4. Homocedasticidade ou variância constante de u_i;

O Método dos Mínimos Quadrados Ordinários

Hipóteses do Método

5. Não existe autocorrelação entre os termos de erro: $cov(u_i, u_j) = 0$;
6. O número de observações (n = tamanho da amostra) deve ser maior que o número de parâmetros;
7. Deve haver variabilidade dos valores de X.

O Método dos Mínimos Quadrados Ordinários

Principais Testes de Consistência do Modelo

- **Alguns Testes de heterocedasticidade, contrário à Homocedasticidade**:

a) Teste de Goldfeld-Quandt – para pequenas amostras;

b) Teste de Breusch-Pagan – para grandes amostras.

- **Teste de autocorrelação dos resíduos**:

a) Teste de Durbin-Watson.

O Método dos Mínimos Quadrados Ordinários

Testes e Estimativas Adicionais

- **Alguns Testes de Normalidade**:

a) Teste de Kolmogorov-Smirnov;

b) Teste de Shapiro-Wilk.

- **Testes de significância dos coeficientes (parâmetros) estimados**:

a) Teste t de Student;

b) Teste F de Snedecor-Fischer;

c) Teste Z (normal).

- **Estimativa dos intervalos de confiança para os parâmetros**

O Método dos Mínimos Quadrados Ordinários

Erros Padrão das Estimativas de (MQO)

$$ep(\hat{\beta}_2) = \frac{\hat{\sigma}}{\sqrt{\sum x_i^2}}$$

$$ep(\hat{\beta}_1) = \sqrt{\frac{\sum x_i^2}{n \sum x_i^2}}\,\hat{\sigma}$$

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-k}}$$

k = número de parâmetros estimados.

O Método dos Mínimos Quadrados Ordinários

Propriedades dos Estimadores de MQO: Teorema de Gauss-Markov

- Cada estimador de MQO é BLUE – Best Linear Unbiased Estimator, dado que:

1. É linear, ou seja, uma função linear de uma variável aleatória, tal como a variável dependente Y na função de regressão;
2. É não viesado, isto é, o seu valor médio ou esperado $E(\hat{\beta}_2)$é igual ao verdadeiro valor β_2;
3. Tem variância mínima na classe de todos os estimadores lineares não viesados; um estimador não viesado com a menor variância é conhecido como um estimador eficiente.

O Método dos Mínimos Quadrados Ordinários

ANOVA – Análise de Variância para uma Regressão por MQO

Fonte da Variação	SQ*	gl	MSQ†
Devido à regressão (SQE)	$\sum \hat{y}_i^2 = \hat{\beta}_2^2 \sum x_i^2$	1	$\hat{\beta}_2^2 \sum x_i^2$
Devido aos resíduos (SQR)	$\sum \hat{u}_i^2$	$n-2$	$\frac{\sum u_i^2}{n-2} = \hat{\sigma}^2$
STQ	$\sum y_i^2$	$n-1$	

Fonte de variação	SQ	gl	MSQ	
Devido à regressão (SQE)	95,4255	1	95,4255	$F = \frac{95,4255}{0,8811}$
Devido aos resíduos (SQR)	9,6928	11	0,8811	$= 108,3026$
STQ	105,1183	12		

Em que:

SQ = Soma dos quadrados;

SQR = Soma dos quadrados dos resíduos;

gl = Graus de liberdade;

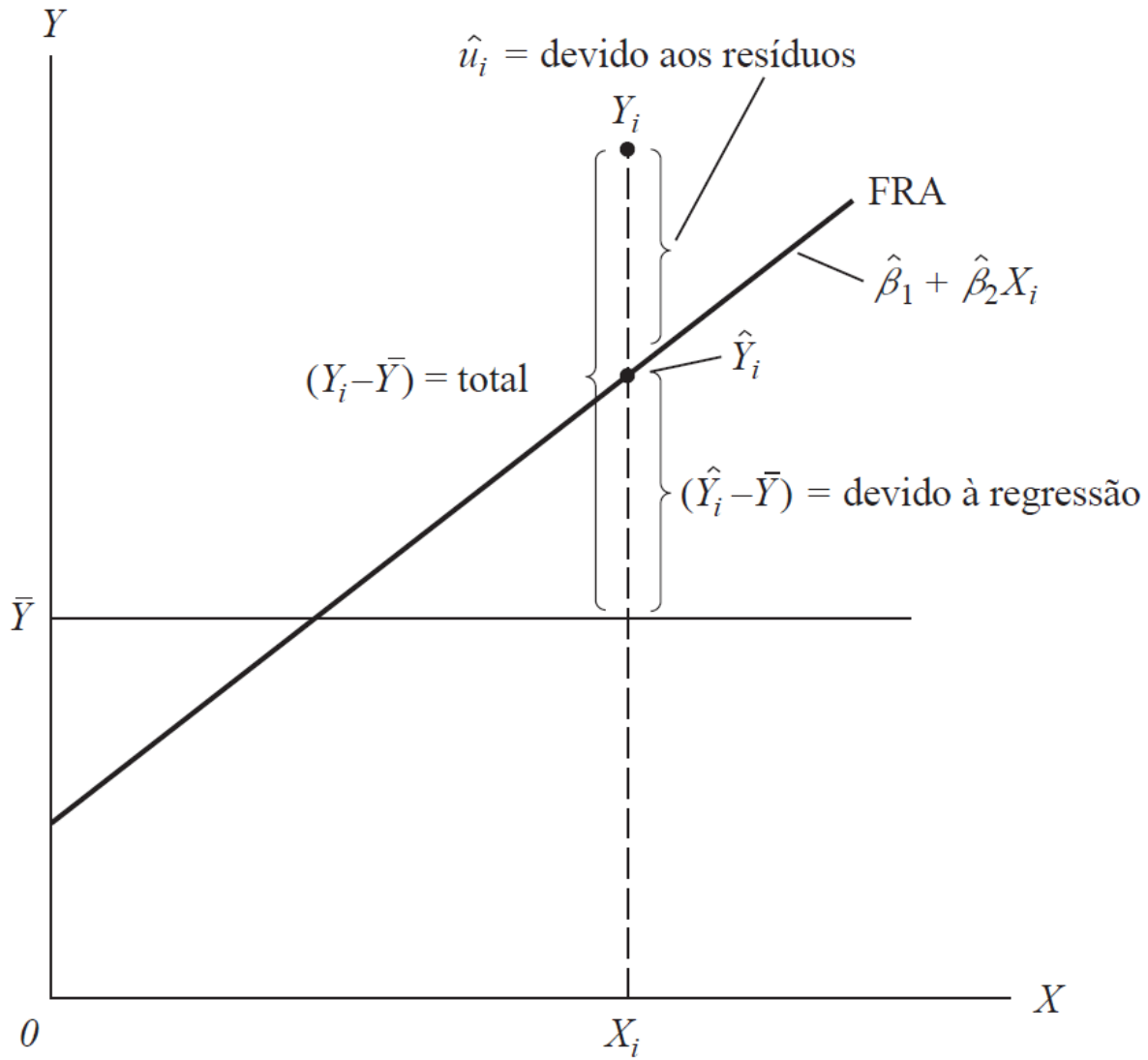
SQE = Soma dos quadrados explicados;

STQ = Soma total dos quadrados;

MSQ = Média da soma dos quadrados.

O Método dos Mínimos Quadrados Ordinários

ANOVA – Análise de Variância para uma Regressão por MQO



O Método dos Mínimos Quadrados Ordinários

Coeficiente de Determinação (R^2)

$$R^2 = \frac{SQE}{SQT} = \frac{\sum \hat{y}_i^2}{\sum y_i^2}$$

➔Medida resumo que diz quanto a reta de regressão amostral se ajusta aos dados, portanto é uma medida da qualidade de ajustamento da reta.

➔Pode-se dizer que o coeficiente de determinação apresenta qual o percentual de explicação da variação total ocorrida na variável dependente, frente as variações das variáveis explicativas.

Por exemplo: R^2 = 0,85 significa que as variáveis explicativas conseguiram explicar 85% do comportamento (variações) da variável dependente.

O Método dos Mínimos Quadrados Ordinários

Coeficiente de Correlação Amostral (R) ou Coeficiente de Correlação Simples ou Coeficiente de Ordem Zero ou Coeficiente Produto Momento de Pearson

$$R = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}} \quad \text{ou} \quad r_{12} = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}$$

Em que:

r_{12}= correlação entre Y (1) e X (2) .

➔No caso de uma regressão simples com duas variáveis (uma variável Y – dependente; e uma variável X – explicativa) significa o grau de associação entre essas duas variáveis. Por exemplo: R = -0,75 significa que quando a vaiável Y cresce em uma unidade, a variável X decresce em média 0,75;

➔Portanto, R pode variar entre: $-1 \leq R \leq 1$;

➔Não é um indicador significantes para descrever relações não lineares;

➔Não implica ou apresenta qualquer relação de causa-efeito;

➔Para uma regressão múltipla (com várias variáveis X – explicativas), o valor de R pode representar a correlação conjunta dos valores de X frente a variável Y, mas é um indicador duvidoso.

O Método dos Mínimos Quadrados Ordinários

Matrizes de Correlação

- **Correlação Parcial ou de primeira ordem**: Indica a correlação entre duas variáveis, mantendo as demais variáveis do modelo constantes.

1. $r_{12,3}$= Coeficiente de correlação parcial entre Y e X_2, mantendo X_3 constante;
2. $r_{13,2}$= Coeficiente de correlação parcial entre Y e X_3, mantendo X_2 constante;
3. $r_{23,1}$= Coeficiente de correlação parcial entre X_2 e X_3, mantendo Y constante.

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \qquad r_{13,2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} \qquad r_{23,1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

O Método dos Mínimos Quadrados Ordinários

A RAZÃO "F"

$$F = \frac{MSQ\ da\ SQE}{MSQ\ da\ SQR}$$

- A Razão F proporciona um teste da hipótese nula H_0: $\beta_s = 0$ ou alternativamente H_a: $\beta_s \neq 0$ estatisticamente.
- Como todas as quantidades que entram nessa equação podem ser obtidas por meio da amostra disponível, essa razão F oferece um teste estatístico para verificar a hipótese nula de que os verdadeiros β_s são estatisticamente iguais a zero.
- Calcula-se a razão F e compara-se com o valor crítico de F (distribuição F) apresentado nas tabelas F ao nível de significância escolhido ou obter o valor p da estatística F calculada.
- Com o teste de F, testa-se a "existência da reta de regressão", pois se todos os betas calculados forem estatisticamente iguais a zero, não existe reta de regressão. Em outras palavras, as variáveis explicativas não explicam nada do comportamento da variável dependente.

Análise de Regressão Linear

Testes de Outliers (observações atípicas)

➔Testes formais (numéricos de maior precisão)

a) Diferem no rigor do teste

b) Metodologias diferentes

➔ Testes visuais (gráficos de menor precisão)

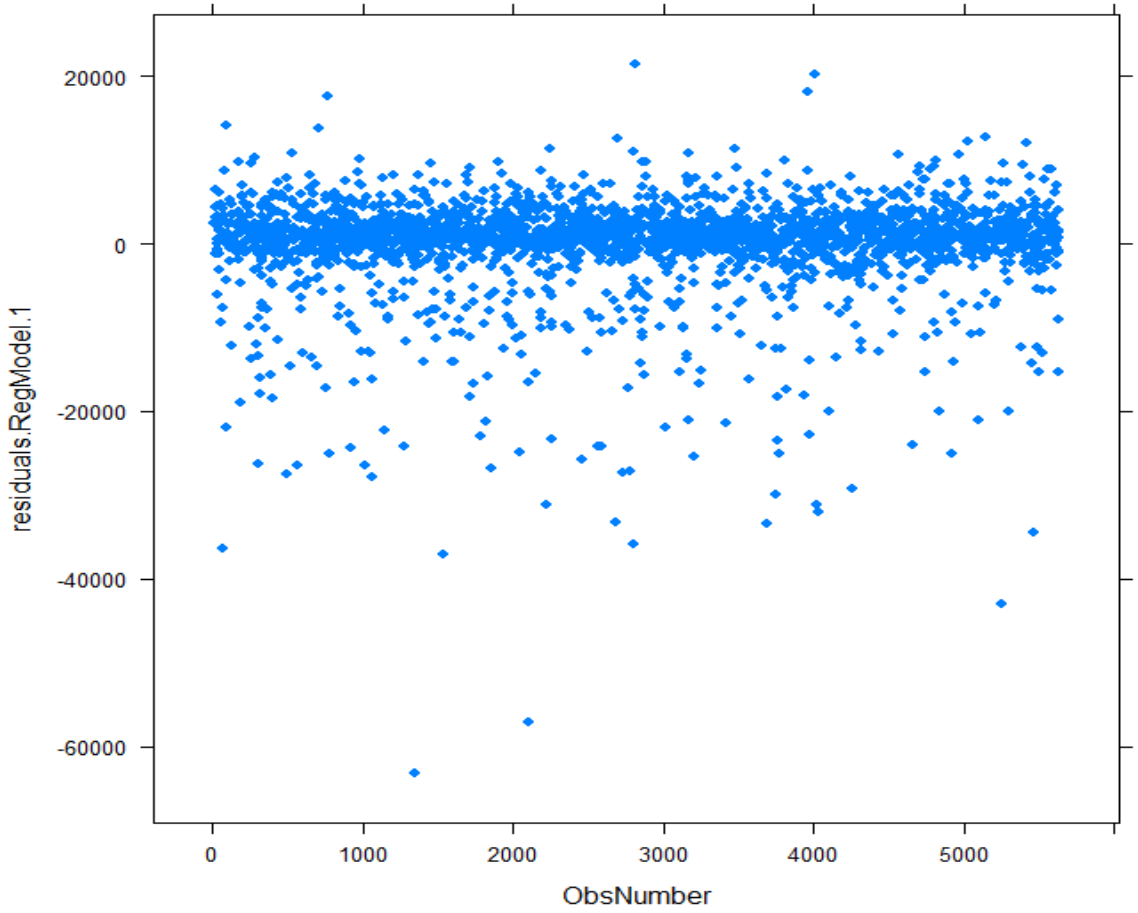
➔A presença de outliers altera significativamente o valor dos parâmetros estimados no modelo, bem como o valor do desvio padrão. Isso pode gerar inclusive o problema da heterocedasticidade.

➔Portanto, se ignorarmos os outliers, os resultados da regressão podem estar equivocados, mas outliers são “ERROS”.

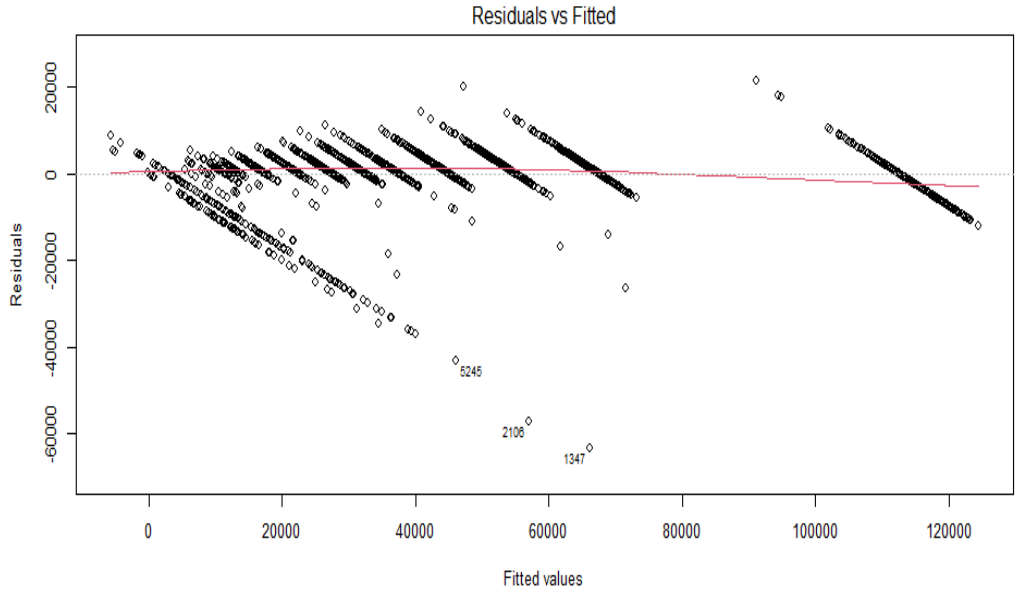
Análise de Regressão Linear

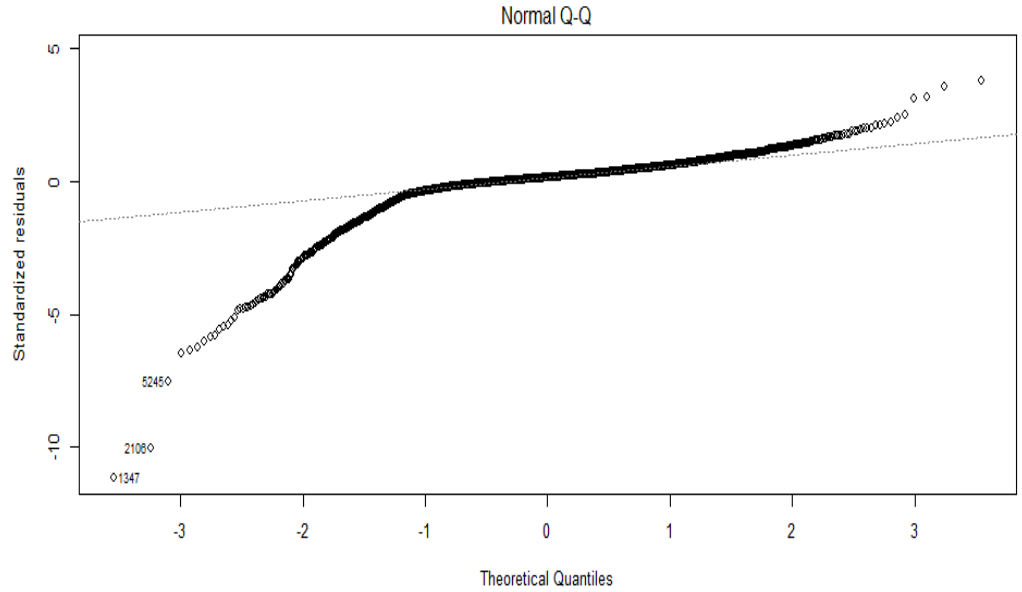
Testes de Outliers (observações atípicas)

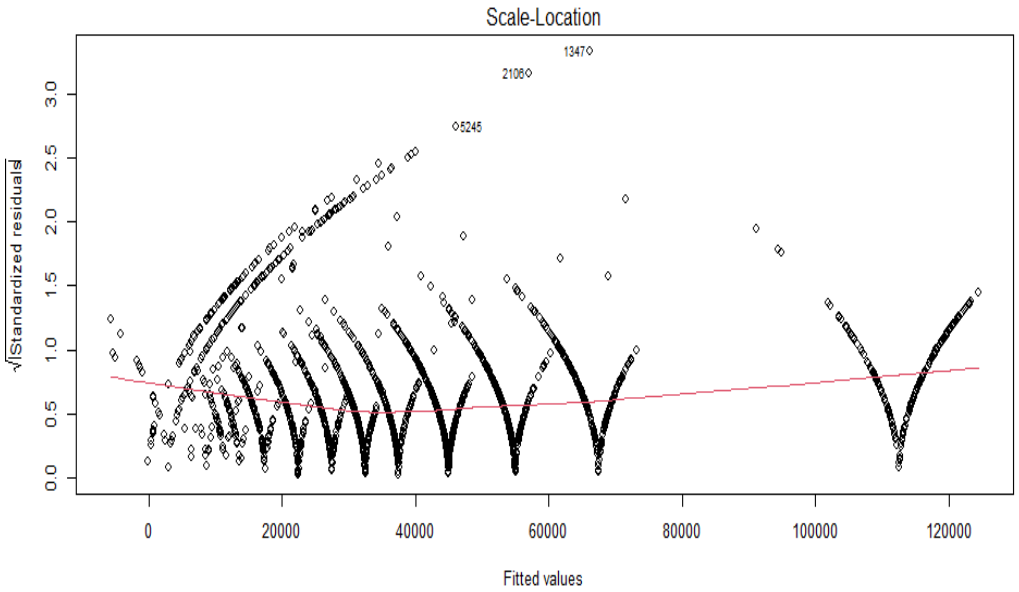
➔Teste visual (pouco preciso)

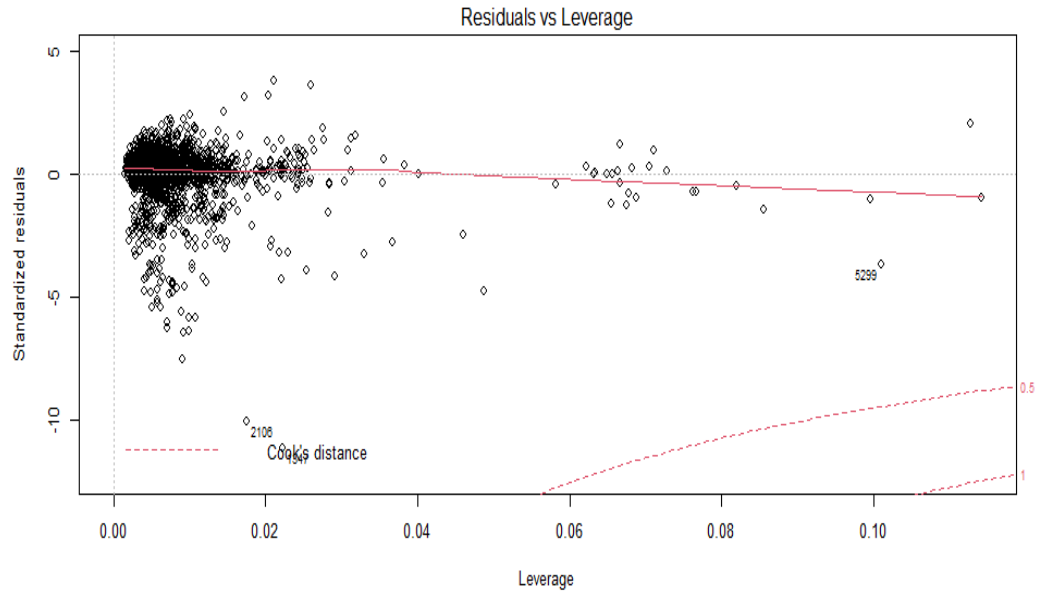


lm(faminc ~ age + black + earns + educ + exper + hispanic + hours + hrwage ...









Análise de Regressão Linear

Testes de Outliers (observações atípicas)

- O gráfico no canto superior esquerdo mostra os erros residuais plotados versus seus valores ajustados. Os resíduos devem ser distribuídos aleatoriamente em torno da linha horizontal, representando um erro residual de zero; isto é, não deve haver uma tendência distinta na distribuição de pontos.
- O gráfico no canto superior direito é um gráfico QQ padrão, o que sugere que os erros residuais são normalmente distribuídos.
- O gráfico de localização da escala no canto inferior esquerdo mostra a raiz quadrada dos resíduos padronizados (tipo de raiz quadrada de erro relativo) em função dos valores ajustados. Novamente, não deve haver tendência óbvia.
- O gráfico no canto inferior direito mostra a alavancagem de cada ponto, que é uma medida de sua importância na determinação do resultado da regressão. Sobrepostas ao gráfico estão linhas de contorno para a distância de Cook, que é outra medida da importância de cada observação para a regressão. Distâncias menores significam que a remoção da observação afeta pouco os resultados da regressão. Distâncias maiores que 1 são suspeitas e sugerem a presença de um possível outlier ou modelo ruim.

Análise de Regressão Linear

Testes de Outliers (observações atípicas)

➔Testes numéricos:

a) Teste de Bonferroni;

b) Distância de Cook;

c) Teste de Dixon;

d) Teste de Grubbs;

e) Outros testes.

Análise de Regressão Linear

Testes de Outliers (observações atípicas)

➔Teste de Bonferroni:

- Reporta os p-values de Bonferroni de cada observação da amostra, através do desvio médio, com base nos resíduos padronizados em modelos lineares (testes t), modelos lineares generalizados (testes de normalidade) e modelos lineares mistos.

- Outros testes veremos na próxima aula.

Análise de Regressão Linear

Correção de Outliers

➔ Deletar as observações que são outliers;

➔ Métodos de suavização (parte deles não garante a resolução do problema);

➔ Não fazer nada – em amostras muito grandes a existência de poucos outliers não influencia nos resultados da regressão.

Análise de Regressão Linear

Teste RESET de Especificação do Modelo

O teste RESET é um diagnóstico popular para correção da forma funcional. A suposição básica é que o modelo pode ser escrito na forma $y = X\beta + Z\gamma$. O "Z" é gerado tomando as potências da resposta ajustada das variáveis dos regressores ou do primeiro componente principal de X. Um teste F padrão é então aplicado para determinar se essas variáveis têm influência adicional significativa. A estatística de teste em H0 segue uma distribuição F com graus de liberdade dos parâmetros.

➔Se o modelo está incorretamente especificado o poder explicativo do modelo ajustado pode ser baixo. Além disso, a relação equivocada entre as variáveis produz resultados espúrios.

➔**Essencial consultar a teoria sobre o fenômeno estudado e outros estudos sobre o mesmo tema.**

Análise de Regressão Linear

Autocorrelação nos Resíduos → Para séries temporais

➔Autocorrelação nos resíduos quer dizer que as observações estão relacionadas entre si. Isto fere um dos pressupostos básicos do modelo.

Consequências:

a) A variância da estimativa é superestimada;

b) A variância dos parâmetros é superestimada;

c) Os parâmetros estimados (betas) não são eficientes (ou seja, são viesados);

d) Superestimativa do R^2;

e) Testes de t e F não são válidos.

Análise de Regressão Linear

Teste de Durbin-Watson para Autocorrelação nos Resíduos

$$d = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n}\hat{u}_t^2}$$

Em que:

d = estatística "d" de Durbin-Watson;

$\hat{u}_t$ = resíduo estimado em "t";

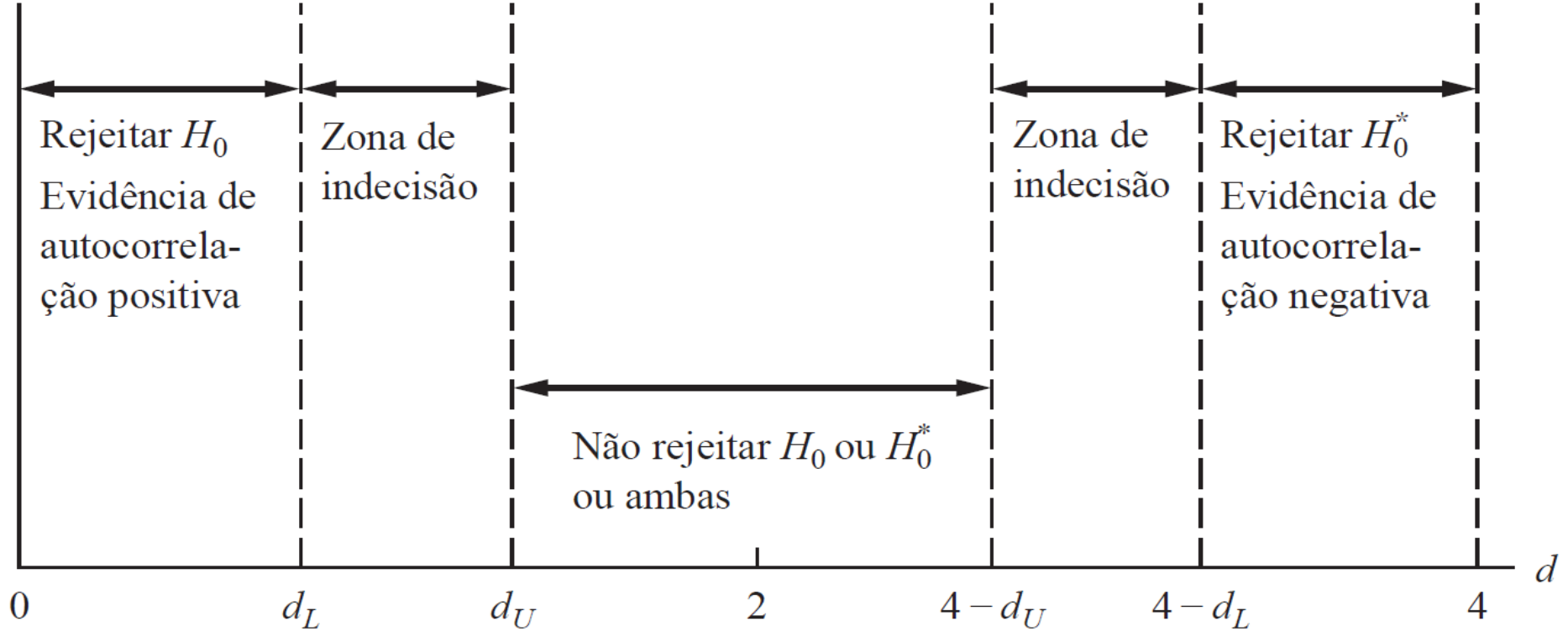
$\hat{u}_{t-1}$ = resíduo estimado em "$t-1$".

Obs: Confrontar o valor calculado de "d" com a tabela de Dubin-Watson, segundo o número de parâmetros e tamanho da amostra.

Análise de Regressão Linear

Teste de Durbin-Watson para Autocorrelação nos Resíduos

Regra de Decisão do Teste



Legenda

H_0: Ausência de autocorrelação positiva

H_0^*: Ausência de autocorrelação negativa

Análise de Regressão Linear

Correção da Autocorrelação nos Resíduos

Capítulo 12 – Gujarati & Porter

- ➔Modelos em primeira diferença;
- ➔Modelos baseados no ρ, baseado na estatística "d" de Durbin-Watson;
- ➔Método de Cochrane-Orcutt;
- ➔Estimadores Sandwich HAC (texto a parte);
- ➔Outras formas.

Análise de Regressão Linear

Homocedasticidade (variância constante)

➔Um dos pressupostos dos MQO é que a variância deve ser constante na amostra. Se não for constante temos o caso de heterocedasticidade. Um erro de especificação do modelo pode gerar heterocedasticidade, bem como a presença de outliers pode ocasionar heterocedasticidade, ou ela pode ser inerente aos dados/fenômeno estudado.

Consequências:

a) Os estimadores (betas) são tendenciosos e ineficientes;

b) Os erros-padrão dos estimadores são superestimados;

c) Os testes de t e F não são válidos;

Análise de Regressão Linear

Teste de Heterocedasticidade de Breusch-Pagan-Godfrey

1) Calcular a variância da estimativa: $\hat{\sigma} = \frac{\hat{u}_i^2}{n-k}$

2) Calcular a proporção do quadrado de cada resíduo em relação a variância da estimativa: $p_i = \frac{\hat{u}_i^2}{\hat{\sigma}^2}$

3) Estimar uma função explicativa desta proporção contra as variáveis explicativas do modelo: $p_i = \alpha_1 + \alpha_2 Z_{2i} + v_i$

4) Calcular: $\theta = \frac{1}{2}(SQE)$

5) Testar o valor calculado de "θ" contra o valor tabelado de qui-quadrado para m-1 graus de liberdade (número de parâmetros -1) : $\theta \sim \chi^2_{m-1}$

Análise de Regressão Linear

Correção da Heterocedasticidade

Capítulo 11 – Gujarati e Porter

- ➔Estimativa por Mínimos Quadrados Ponderados (Regressão Robusta);
- ➔Estimativa por estimadores Sandwich (texto a parte);
- ➔Erros padrão robustos de White

Análise de Regressão Linear

A Escolha do Melhor Modelo Estatístico

➔Nem todas as variáveis explicativas que incluímos em modelos estatísticos são significativas e explicam o comportamento da variável dependente que estamos estudando.

➔Para inferência estatística é recomendável a definição de um modelo que inclua apenas as variáveis com poder explicativo significativo, ao nível de significância escolhido (geralmente 5% - 95% de confiança).

➔A escolha do melhor modelo que representa um determinado fenômeno é muito importante, especialmente se desejamos fazer inferência estatística.

Análise de Regressão Linear

A Escolha do Melhor Modelo Estatístico

➔Os indicadores para a escolha do melhor modelo geralmente são: R^2, R^2 ajustado, AIC, BIC, AICc.

➔ Esses são indicadores consideram:

a) Grau de ajustamento da amostra à reta de regressão estimada;

b) Tamanho da amostra;

c) Número de variáveis constantes no modelo.

Análise de Regressão Linear

A Escolha do Melhor Modelo Estatístico

O R^2 como indicador

$$R^2 = \frac{SQE}{STQ}$$

em que:

R^2 = Coeficiente de determinação;

SQE = Soma dos quadrados explicados;

STQ = Soma total dos quadrados.

Obs: O valor se situa entre 0 e 1, um valor de 0.89, significa que 89% das variações na variável dependente (Y) foram explicadas pelas variáveis explicativas do modelo.

Em uma função de regressão com 2 variáveis X (X_2 e X_3)tem-se:

$$R^2 = \frac{\hat{\beta}_2 \sum y_i x_{2i} + \hat{\beta}_3 \sum y_i x_{3i}}{\sum y_i^2} = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} = \frac{SQE}{STQ}$$

Análise de Regressão Linear

A Escolha do Melhor Modelo Estatístico

O R^2 Ajustado como indicador

$$R^2adj = 1 - \frac{\sum \hat{u}_i^2/(n-k)}{\sum y_i^2/(n-1)}$$

em que:

$R^2\ adj$ = Coeficiente de determinação ajustado;

$\sum \hat{u}_i^2$ = Somatória dos resíduos ao quadrado;

$\sum y_i^2$= Somatória dos valores de “y” ao quadrado;

n = Tamanho da amostra;

k= Número de parâmetros do modelo.

Obs: O valor se situa entre 0 e 1, um valor de 0.71, significa que 71% das variações na variável dependente (Y) foram explicadas pelas variáveis explicativas do modelo, ponderados o tamanho da amostra e o número de variáveis incluídas no modelo.

Análise de Regressão Linear

A Escolha do Melhor Modelo Estatístico

O AIC como indicador

$$AIC = -2\log(\hat{\theta}) + 2p$$

em que:

AIC = Critério de Informação de Akaike;

$\hat{\theta}$= Função de máxima verossimilhança do modelo;

p = Número de variáveis explicativas do modelo.

Obs: quanto menor o valor AIC, melhor é o modelo.

Análise de Regressão Linear

A Escolha do Melhor Modelo Estatístico

O BIC como indicador

$$BIC = -2 \log f(x_n|\theta) + p.\log(n)$$

em que:

BIC = Critério de Informação Bayesiano;

$f(x_n|\theta)$ = Função de máxima verossimilhança do modelo escolhido;

p = Número de parâmetros;

n = Tamanho da amostra.

Obs: quanto menor o valor BIC, melhor é o modelo.

Análise de Regressão Linear

A Escolha do Melhor Modelo Estatístico

O AICc como indicador

$$AICc = -2\log\left(\hat{\theta}\right) + 2p + 2\frac{p(p+1)}{n-p-1}$$

em que:

$AICc$ = Critério de Informação Bayesiano;

$\hat{\theta}$= Função de máxima verossimilhança do modelo;

p = Número de variáveis explicativas do modelo;

n = Tamanho da amostra.

Obs: quanto menor o valor AICc, melhor é o modelo.

Análise de Regressão Linear

A Escolha do Melhor Modelo Estatístico – Regressão Stepwise

➔A regressão Stepwise é uma forma de, por exclusão e inclusão de variáveis explicativas no modelo estatístico, estimar o modelo de melhor desempenho.

➔O Stepwise pode ser: Backward; Forward; ou Backward/Forward

➔No Backward Stepwise considera-se inicialmente o modelo com todas as variáveis e o procedimento extrai uma variável não significativa (aquela de maior p-value) de cada vez e repete nova regressão. Ao final a escolha do melhor modelo é dado comparando os valores AIC, BIC ou AICc de cada modelo estimado.

Análise de Regressão Linear

A Escolha do Melhor Modelo Estatístico – Regressão Stepwise

➔No Forward Stepwise inicia-se com um modelo mínimo, de apenas uma variável explicativa (aquela que tem maior poder explicativo, comparando as variáveis X), após isso vai-se adicionando uma nova variável ao modelo, que é reestimado. Ao final a escolha do melhor modelo é dado comparando os valores AIC, BIC ou AICc de cada modelo estimado.

➔O Backward/Forward Stepwise é a junção dos métodos Backward e Forward para escolher o melhor modelo.

Análise de regressão – Introd.

Coefficients:

		Estimate	Std. Error	t	value Pr(>|t|)
(Intercept)	β_0	6.349e+02	4.027e+02	1.577	0.11673
age	β_1	6.125e+00	5.552e+00	1.103	0.27146
college	β_2	-1.527e+02	2.520e+02	-0.606	0.54541
comten	β_3	-3.783e+00	3.872e+00	-0.977	0.33000
grad	β_4	-5.935e+01	8.540e+01	-0.695	0.48805
mktval	β_5	2.664e-02	9.717e-03	2.741	0.00677 **
sales	β_6	1.543e-02	1.019e-02	1.514	0.13195

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 538.2 on 170 degrees of freedom

Multiple R-squared: 0.1896, Adjusted R-squared: 0.161

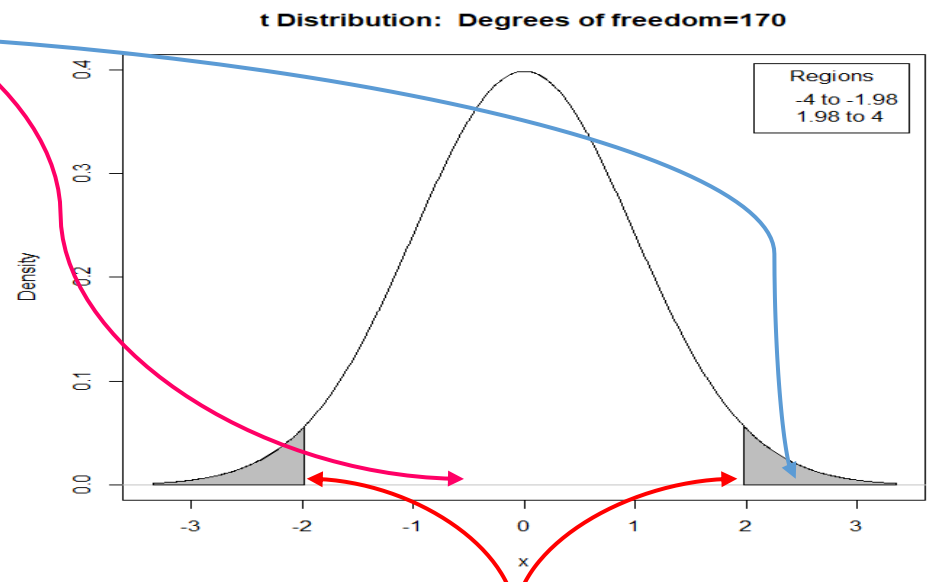
F-statistic: 6.63 on 6 and 170 DF, p-value: 2.569e-06

Obs: deve-se fazer teste de t para todos os coeficientes calculados

Exemplos de testes de hipótese

H0: $\beta_4 = 0$
Ha: $\beta_4 \neq 0$

H0: $\beta_5 = 0$
Ha: $\beta_5 \neq 0$



t_{tab} = -1.98 | 1.98

Análise de regressão – Introd.

Residual standard error: 538.2 on 170 degrees of freedom

Multiple R-squared: 0.1896, Adjusted R-squared: 0.161

F-statistic: 6.63 on 6 and 170 DF, p-value: 2.569e-06

R-quadrado = 0.1896 ➔ quer dizer que as variáveis explicativas conseguem explicar 18,96% das variações da variável dependente. Ou seja, essas variáveis explicativas incluídas no modelo conseguem explicar somente 18,96% do salário dos CEOs. O R-quadrado considera apenas as variações ocorridas nas variáveis.

R-quadrado ajustado = 0.161 ➔ quer dizer que as variáveis explicativas conseguem explicar 16,10% das variações ocorridas nos salários dos CEOs. Este indicador considera as variações ocorridas nas variáveis, bem como o tamanho da amostra e o número de variáveis do modelo. Em outras palavras indica com maior precisão se o modelo é parcimonioso.

➔ Esse dois indicadores acima representam são usualmente chamados de coeficiente de determinação e apresentam a qualidade de ajustamento da reta de regressão aos dados das variáveis utilizadas no modelo.

Análise de regressão – Introd.

Residual standard error: 538.2 on 170 degrees of freedom

Multiple R-squared: 0.1896, Adjusted R-squared: 0.161

F-statistic: 6.63 on 6 and 170 DF, p-value: 2.569e-06

A estatística F testa se todos os parâmetros são estatisticamente diferentes de "zero", ou seja se existe reta de regressão.

$H0: \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 = 0$

$Ha: \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 \neq 0$

Resultado do teste:

Como a estatística calculada situa-se na área de rejeição, rejeita-se H0 de que todos os coeficientes conjuntamente são iguais a "zero", em favor da hipótese alternativa de que pelo menos um dos parâmetros calculados é diferente de "zero", ou seja, existe reta de regressão.