Estatística para Ciências de Dados

Aula 7: Modelos de Regressão Linear

Mariana Cúri
ICMC/USP
mcuri@icmc.usp.br

Conteúdo

1. Introdução
    a. Modelo
    b. Interpretação
    c. Notação matricial
2. Estimação dos parâmetros
    a. Mínimos Quadrados Ordinários (MQO)
    b. Máxima Verossimilhança (MV)
3. Ajuste do modelo
4. Inferência sobre os parâmetros de regressão
5. Comparação de modelos
6. Tópicos adicionais

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Modelo

1. Objetivos:

- prever Y a partir do conhecimento de **X**=**x**
- verificar a importância de cada variável em **X** na previsão de Y

linear nos parâmetros

2. Quando **X** é composto de uma única variável: Regressão Linear Simples

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \text{ , para } i = 1, \cdots, n$$

3. Se **X** é composto por mais de uma variável: Regressão Linear Múltipla

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon_i$$

Notação matricial: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Modelo e interpretação (regressão linear simples)

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad i = 1, \cdots, n$$

intercepto

variável dependente
ou variável resposta

parâmetros de regressão

erro aleatório

variável independente
ou variável explicativa
ou variável preditora (**fixada**)

i-ésima unidade
da amostra

SUPOSIÇÕES

$$E(\epsilon_i) = 0$$

$$V(\epsilon_i) = \sigma^2$$

ϵ_i , ϵ_j : não correlacionados para i ≠ j

CONSEQUÊNCIAS

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i + E(\epsilon_i) \\ &= \beta_0 + \beta_1 X_i \end{aligned}$$

$$V(Y_i) = \sigma^2$$

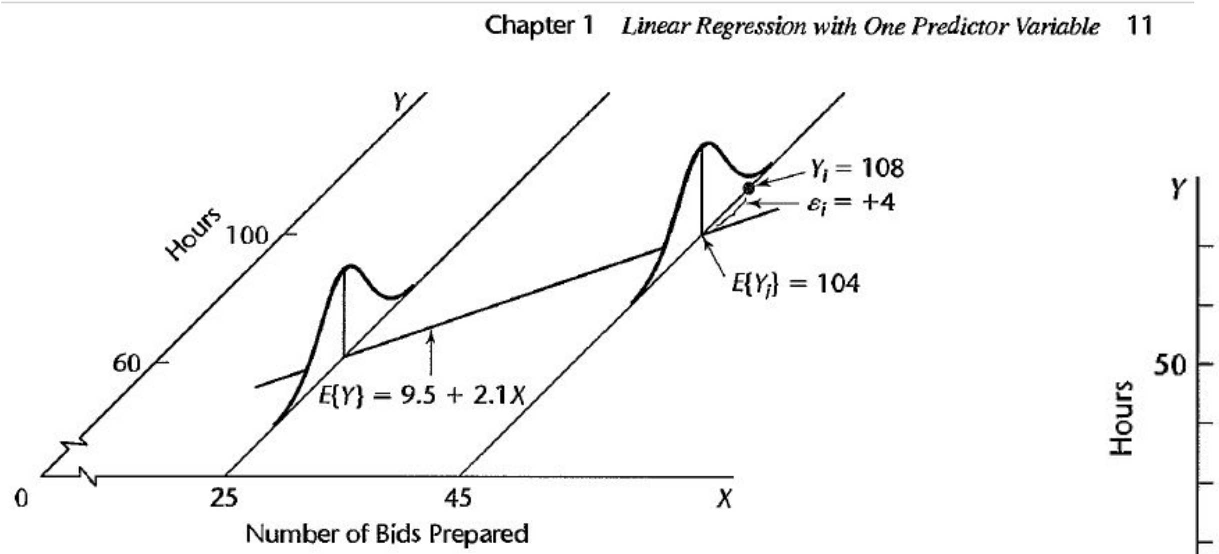
Y_i, Y_j : não correlacionados para i ≠ j

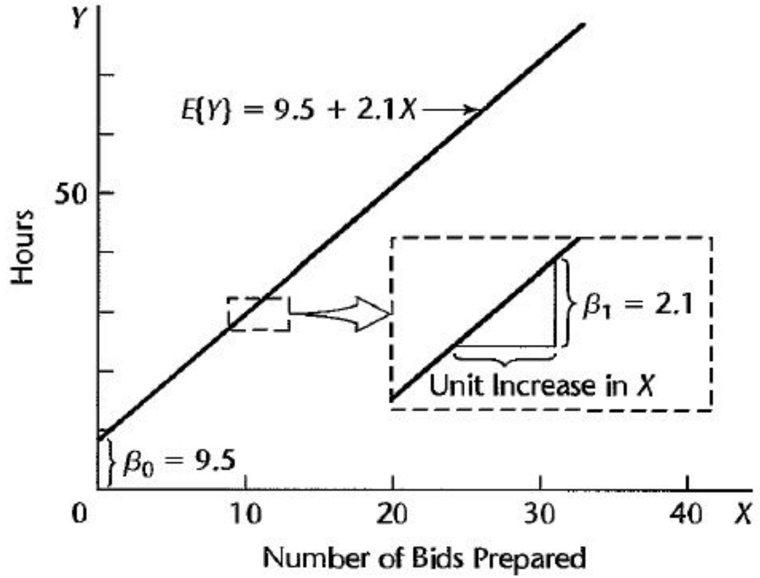
β_0 : E(Y) para X=0

β_1 : quanto varia E(Y) para o aumento de 1 unidade em X

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida

Modelo e interpretação (regressão linear simples)





Fonte: Kutner et al. Applied Linear Statistical Models

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Modelo centrado (regressão linear simples)

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \epsilon_i$$

$$i = 1, \cdots, n$$

SUPOSIÇÕES

$$E(\epsilon_i) = 0$$

$$V(\epsilon_i) = \sigma^2$$

ϵ_i , ϵ_j : não correlacionados para i ≠ j

CONSEQUÊNCIAS

$$E(Y_i) = \beta_0^* + \beta_1(X_i - \bar{X})$$

$$V(Y_i) = \sigma^2$$

Y_i, Y_j : não correlacionados para i ≠ j

β_0^* : E(Y) para X igual a média amostral de X

β_1 : quanto varia E(Y) para o aumento de 1 unidade em X

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Notação matricial (regressão linear múltipla)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p-1,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p-1,2} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{p-1,n} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} \qquad V(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}$$

β_0 : E(Y) para $X_1 = X_2 = \ldots = X_{p-1} = 0$

β_j : quanto varia E(Y) para o aumento de 1 unidade em X_j, j = 1, ... p-1, e as demais variáveis explicativas mantidas fixas num mesmo valor

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

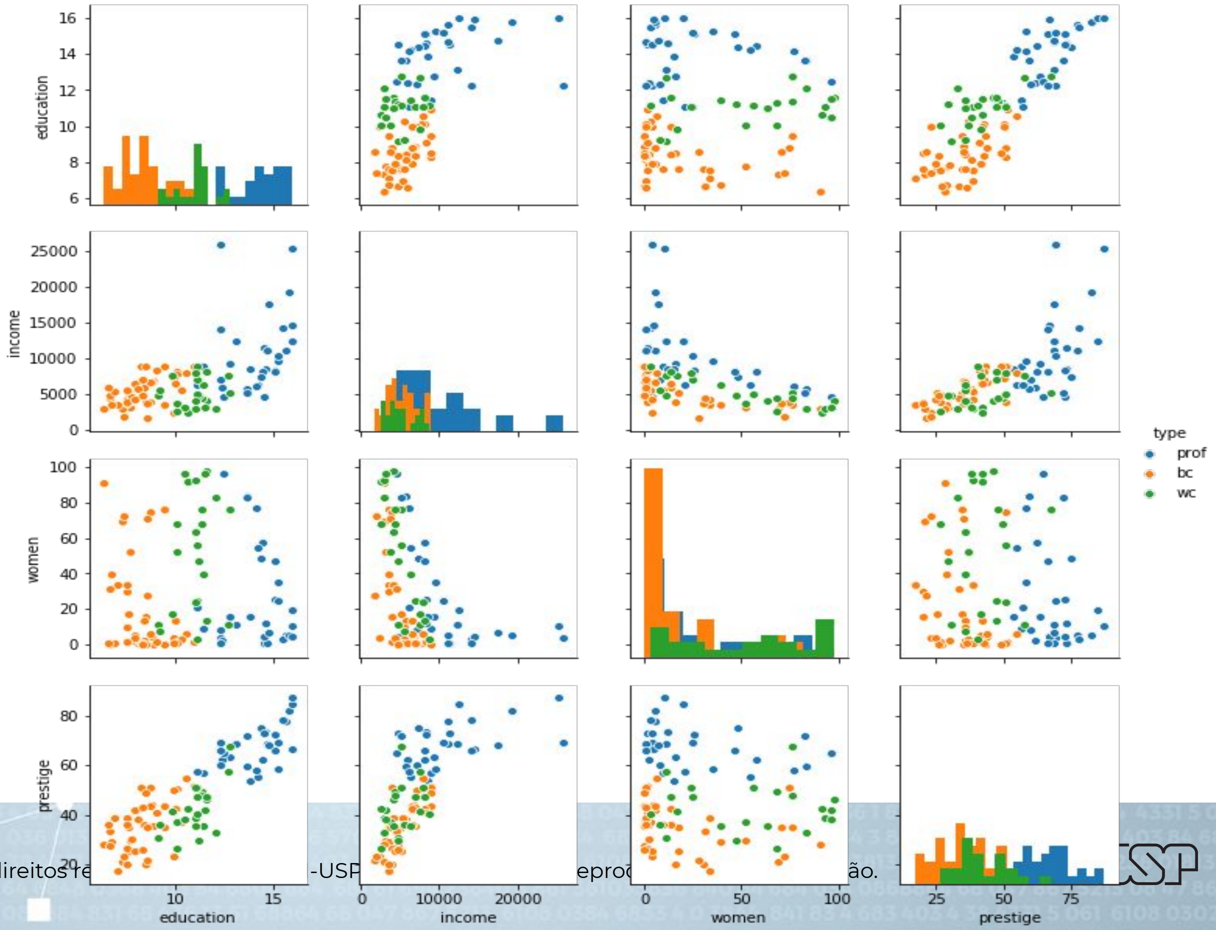
Exemplo: Prestígio ocupação profissional

	X1	X2	X3	Y		X4
occupation	education	income	women	prestige	census	type
gov.administrators	13.11	12351	11.16	68.8	1113	prof
general.managers	12.26	25879	4.02	69.1	1130	prof
accountants	12.77	9271	15.7	63.4	1171	prof
purchasing.officers	11.42	8865	9.11	56.8	1175	prof
chemists	14.62	8403	11.68	73.5	2111	prof
physicists	15.64	11030	5.13	77.6	2113	prof
biologists						
architects						
civil.engineers						
mining.engineers						
surveyors						
draughtsmen						
computer.programers						
economists						
psychologists						
social.workers						
lawyers	15.77	19263	5.13	82.3	2343	prof
librarians	14.15	6112	77.1	58.1	2351	prof
vocational.counsellors	15.22	9593	34.89	58.3	2391	prof
ministers	14.5	4686	4.14	72.8	2511	prof

Average education of occupational incumbents, years
Average income of incumbents, dollars
Percentage of incumbents who are women
Pineo-Porter prestige score for occupation
Canadian Census occupational code
Type of occupation
(bc: Blue Collar; prof: Professional, Managerial, and Technical; wc: White Collar)

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Exemplo: Prestígio ocupação profissional



Copyright © 2020. Todos os direitos re... -USP ... reprod... ão.

Estimação

MQO, minimizar:

$$Q = \sum_{i=1}^{n} (Y_i - E(Y_i))^2$$

$$= \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}))^2$$

$$= (\boldsymbol{Y} - \boldsymbol{X\beta})^t (\boldsymbol{Y} - \boldsymbol{X\beta})$$

MV, maximizar:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}))^2}$$

$$\hat{\sigma}^2 = MSE = \frac{SQE}{n-p}$$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{Y}$$

Propriedades:
- não viciado de variância mínima
- consistentes
- suficientes

Estimador da E(Y)

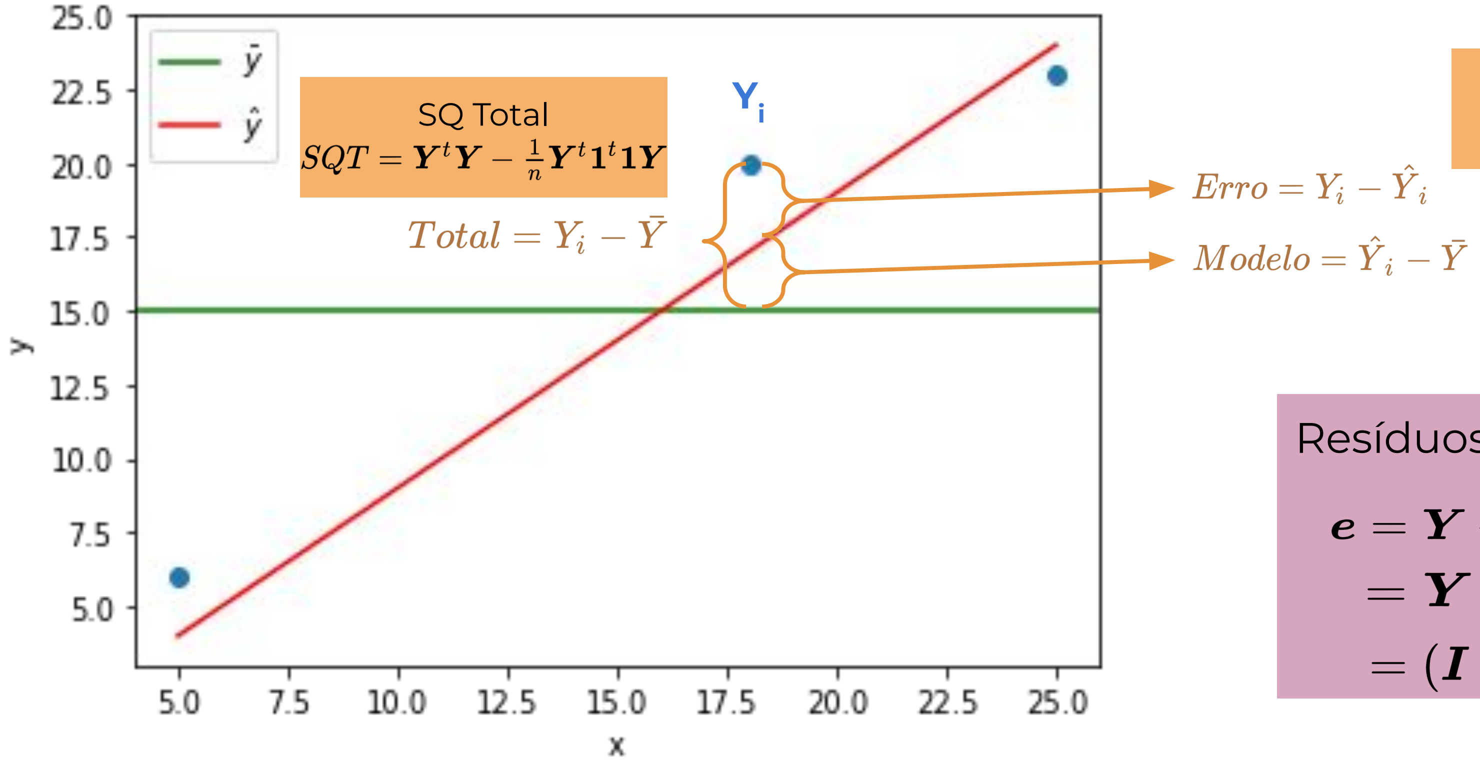
$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

$$= \boldsymbol{X}(\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{Y}$$

$$= \boldsymbol{HY}$$

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Estimação



SQ dos Resíduos

$$SQE = \boldsymbol{e}^t\boldsymbol{e}$$

Resíduos (ou Erro)

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{Y} - \hat{\boldsymbol{Y}} \\ &= \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} \\ &= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \end{aligned}$$

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Ajuste do modelo

Coeficiente de Determinação (ou de Explicação):

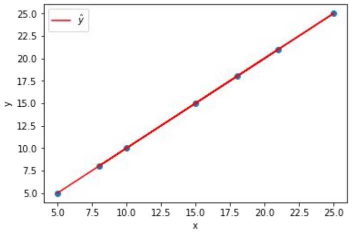
$$R^2 = 1 - \frac{SQE}{SQT}$$

$$0 \leq R^2 \leq 1$$

qual o % da variabilidade de Y explicada pelo modelo

nunca diminui, se variáveis explicativas são acrescentadas

$$R^2 = 1$$



Coeficiente de Determinação Ajustado (pelo nº de variáveis explicativas):

$$R_a^2 = 1 - \frac{n-1}{n-p}\frac{SQE}{SQT}$$

$$0 \leq R_a^2 \leq 1$$

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Ajuste do modelo: resíduos

Análise dos resíduos para detecção de *outliers* em Y

Resíduo padronizado: $$\frac{e_i}{\sqrt{MSE}}$$

Resíduo studentizado (internamente): $$\frac{e_i}{\sqrt{Var(e_i)}} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

Resíduo studentizado externamente: exclui a i-ésima unidade amostral para o cálculo do valor predito

$$\frac{y_i - \hat{y}_{i(i)}}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Ajuste do modelo: resíduos

Análise dos resíduos para detecção de *outliers* em x

Alavanca do i-ésimo caso: medida de distância entre X_i e $\bar{X}$

h_{ii}

$0 \leq h_{ii} \leq 1$

$\sum_{i=1}^{n} h_{ii} = p$

$$h_{ii} > 2\bar{h} = 2p/n$$

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Ajuste do modelo: resíduos

Análise dos pontos influentes: (nas estimativas dos valores preditos)

DFFITS: medida da influência de Y_i em $\hat{Y}_i$

$$\frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MSE_{(i)}(1-h_{ii})}}$$

> 1 ou $> 2\sqrt{p/n}$

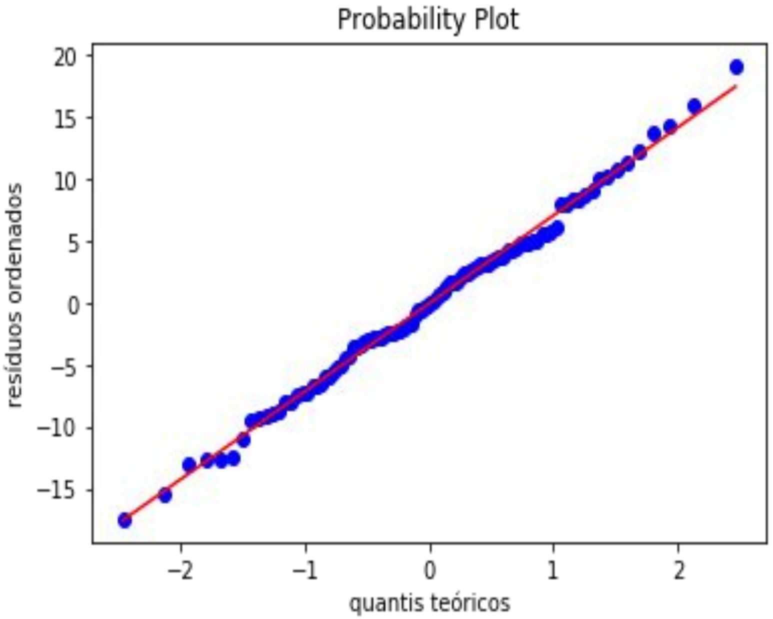
Distância de Cook: medida da influência de Y_i em todos os valores ajustados

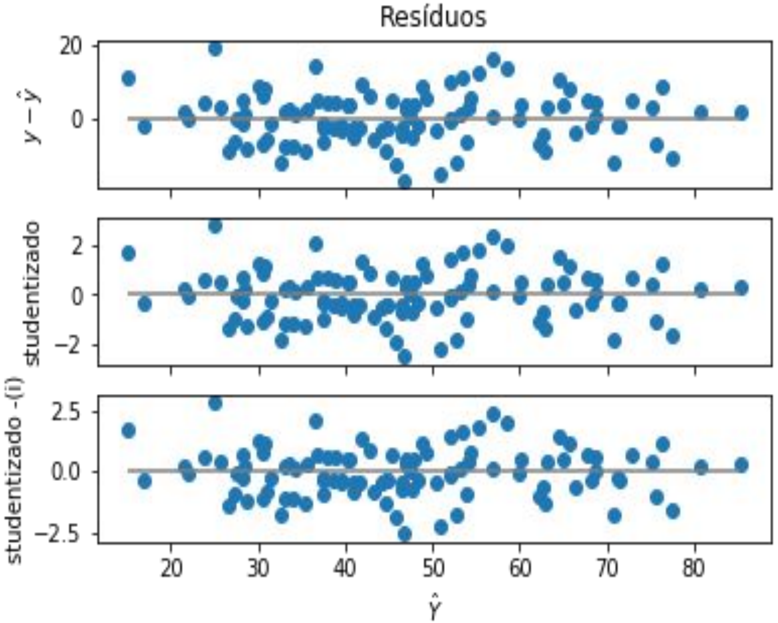
$$\frac{\sum_{i=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{pMSE}$$

comparar com percentil da F(p,n-p)

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

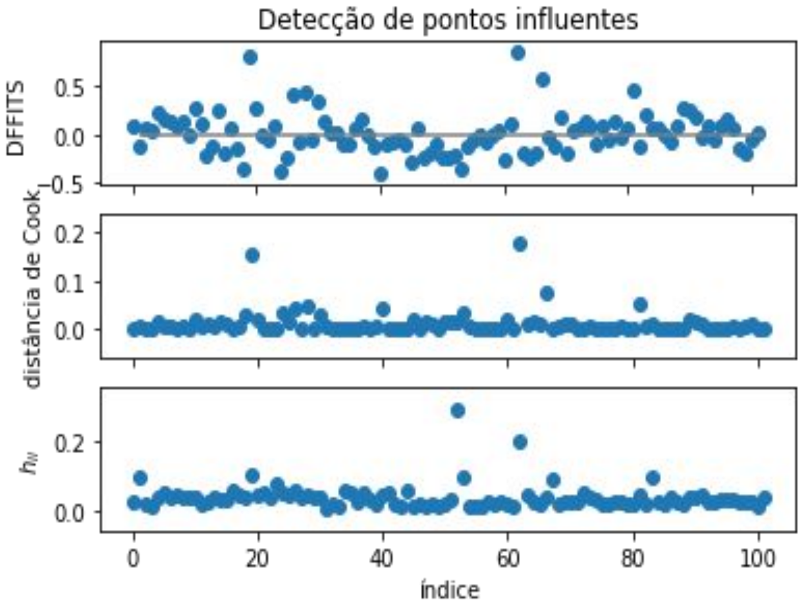
Ajuste do modelo: resíduos

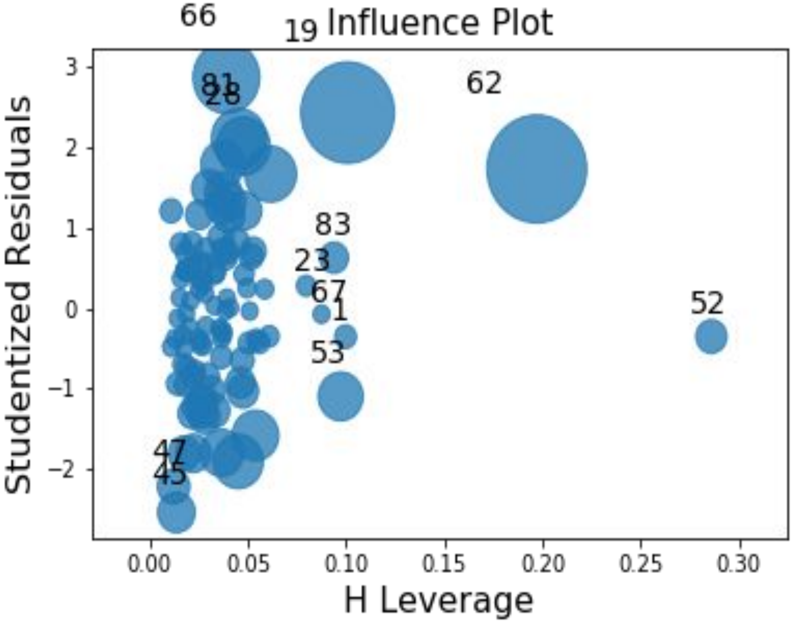




Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Ajuste do modelo: resíduos





Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Inferência sobre os parâmetros de regressão

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t \boldsymbol{X})^{-1} \boldsymbol{X}^t \boldsymbol{Y}$$

Estimador da variância de $\hat{\boldsymbol{\beta}}$: $\boldsymbol{s}^2(\hat{\boldsymbol{\beta}}) = MSE(\boldsymbol{X}^t \boldsymbol{X})^{-1}$

Sob distribuição Normal dos erros aleatórios: $\frac{\hat{\beta}_k - \beta_k}{\boldsymbol{s}(\hat{\beta}_k)} \sim t_{n-p}$

Suficiente para construir IC e testar hipóteses sobre os parâmetros

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Inferência sobre os parâmetros de regressão

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            df.prestige   R-squared:                       0.835
Model:                            OLS   Adj. R-squared:                  0.830
Method:                 Least Squares   F-statistic:                     165.4
Date:                Thu, 21 May 2020   Prob (F-statistic):           3.21e-38
Time:                        17:02:36   Log-Likelihood:                -342.51
No. Observations:                 102   AIC:                             693.0
Df Residuals:                      98   BIC:                             703.5
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept     -110.9658     14.843     -7.476      0.000    -140.421     -81.511
l_income         9.3147      1.327      7.022      0.000       6.682      11.947
df.education     3.7305      0.354     10.527      0.000       3.027       4.434
df.women         0.0469      0.030      1.568      0.120      -0.012       0.106
==============================================================================
```

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Inferência sobre os parâmetros de regressão

Effects of Departures from Normality

If the probability distributions of Y are not exactly normal but do not depart seriously, the sampling distributions of b_0 and b_1 will be approximately normal, and the use of the t distribution will provide approximately the specified confidence coefficient or level of significance. Even if the distributions of Y are far from normal, the estimators b_0 and b_1 generally have the property of *asymptotic normality*—their distributions approach normality under very general conditions as the sample size increases. Thus, with sufficiently large samples, the confidence intervals and decision rules given earlier still apply even if the probability distributions of Y depart far from normality. For large samples, the t value is, of course, replaced by the z value for the standard normal distribution.

Fonte: Kutner et al. Applied Linear Statistical Models

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Comparação de modelos

1. Critérios para seleção de modelos: todos os 2^{p-1} modelos possíveis
   a. R_a^2
   b. MSE
   c. AIC = n *ln*(SQE) - n *ln*(n) + 2p
   d. Schwarz' BIC = n *ln*(SQE) - n *ln*(n) + *ln*(n) p
2. Seleção automática: muitas variáveis explicativas
   a. <u>Backward</u>: modelo completo, exclui uma variável preditora de cada vez
   b. <u>Forward</u>: modelo apenas com o intercepto, inclui um preditor de cada vez
   c. <u>Stepwise</u>: forward que permite que um preditor que já está no modelo possa sair

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.

Tópicos adicionais

- Variáveis explicativas categorizadas
- Interação entre variáveis explicativas
- Validação: cross-validation, treino-teste
- Multicolinearidade ou muitos preditores (seleção automática)
    - Regressão ridge
    - Regressão LASSO

Copyright © 2020. Todos os direitos reservados ao CeMEAI-USP. Proibida a cópia e reprodução sem autorização.