Modelo de Regressão Múltipla

Modelo de Regressão Linear Simples

- Última aula:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

  - Y é a variável resposta;
  - X é a variável independente;
  - ε representa o erro.

Modelo Clássico de Regressão Múltipla

Modelo clássico de regressão é definido por:

(i) "*n*" respostas y_i independentes, tendo cada y_i uma distribuição especificada de média $\boldsymbol{\mu_i = E(y_i)}$ e **variância $\boldsymbol{\sigma^2}$ constante**.

(ii) a **média** $\boldsymbol{\mu_i}$ é expressa de forma linear por $\boldsymbol{\mu_i = x_i^T\beta}$, onde x_i^T é um vetor *1xp* com valores de *p* variáveis explicativas relacionadas a i-ésima resposta y_i e β é um vetor *px1* de parâmetros a serem estimados.

$$y = \mu + \varepsilon = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}}_{\mu} + \varepsilon$$

Modelo Clássico de Regressão (MCR)

- Em geral adota-se a hipótese de aditividade entre *y* e μ, isto é, ***y* = μ + ε**, onde ε é um vetor de erros com **E(ε) = 0** e **Var(ε) = σ^2**.
- A estimação de β *pode* ser feita pelo Método dos Mínimos Quadrados, que não requer qualquer hipótese sobre a distribuição das componentes do vetor *y*, e consiste em minimizar

$$\sum_{i=1}^{n}(y_i - \mu_i)^2 = \sum_{i=1}^{n}(\varepsilon_i)^2 = SQE(\beta)$$

MCR - Estimação

O modelo clássico de regressão é representado em **notação matricial** por

$$y_{(nx1)} = X_{(nxp)}\beta_{(px1)} + \varepsilon_{(nx1)} \qquad (1)$$

X = matriz modelo, suposta de posto p;

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ 1 & x_{21} & \cdots & x_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

MCR - Estimação

- Para estimar β minimiza-se SQE (β) em relação β.
- Solução de um sistema de *p* equações lineares dadas por

$$\frac{\partial SQE(\beta)}{\partial \beta_r} = 0, \qquad r = 0, \dots, p-1.$$

- Em notação matricial o sistema é expresso por

$$X^T X\beta = X^T y$$

MCR - Estimação

- Essas equações lineares são conhecidas como *equações normais*.
- Se a matriz X tem posto completo, então X^TX é inversível e, portanto, a solução do sistema de equações normais é única.
- A solução corresponde ao *estimador de mínimos quadrados* (EMQ) de β dado por

$$\hat{\beta} = (X^T X)^{-1} X^T y \qquad (2)$$

MCR - Estimação

O EMQ $\hat{\beta}$ em (2), segundo o modelo (1), tem as seguintes propriedades:

(i) $\hat{\beta}$ minimiza $\sum_i \varepsilon_i^2$, **independente** da distribuição proposta para os erros. (entretanto a normalidade será necessária para realizarmos inferência sobre os parâmetros β).

(ii) as componentes do vetor $\hat{\beta}$ são funções lineares das observações e são **estimadores não-viesados de menor variância** dos parâmetros em β.

MCR - Estimação

A *soma de quadrados dos resíduos* (SQR) mede a discrepância entre o vetor *y* e o vetor de valores ajustados $\hat{\mu} = X\hat{\beta}$, sendo expresso por:

$$SQR = SQE(\hat{\beta}) = (y - X\hat{\beta})^T (y - X\hat{\beta})$$

Assim, o vetor de erros não observados $\varepsilon = y - X\beta$ é estimado pelo **vetor de resíduos *r***, dado por

$$r = y - \hat{\mu} = y - X\hat{\beta}. \qquad (3)$$

Propriedades do EMQ e dos Resíduos

As propriedades abaixo são baseadas apenas nas duas hipóteses básicas atribuídas aos erros:

$E(\varepsilon) = 0$ e $Cov(\varepsilon) = \sigma^2 I$.

(a) O EMQ $\hat{\beta}$ é não-viesado. ($E(\hat{\beta}) = \beta$);

(b) A covariância do EMQ $\hat{\beta}$ é expressa por

$$Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \qquad (4)$$

OBS: Os elementos da diagonal são as variâncias das EMQ de β e, portanto representam a precisão das estimativas.

Propriedades do EMQ e dos Resíduos

(c) Um estimador não-viesado de σ^2 é dado por

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n-p} \qquad (5)$$

(d) Esperança e Covariância do vetor de Resíduos

$$E(r) = 0;$$

$$\mathrm{Cov}(r) = \sigma^2(I-H) \Rightarrow \mathrm{Cov}(r_i, r_j) = \sigma^2(1-h_{ij}).$$

OBS: Assim, embora os erros aleatórios ε_i sejam homocedásticos, o mesmo não ocorrem com os resíduos, cujas variâncias dependem dos elementos da diagonal da matriz de projeção $H = X(X^TX)^{-1}X^T$.

Modelo Normal Linear

- A especificação de uma distribuição para os erros aleatórios faz-se necessária para determinarmos a distribuição de probabilidade das EMQ.
- A suposição de normalidade dos erros é a mais adotada e considera que $\varepsilon_1, \ldots, \varepsilon_n$ em (1) são independentes e tem distribuição normal $N(0, \sigma^2)$.

Modelo Normal Linear

Segundo a hipótese de normalidade dos erros, temos as seguintes propriedades:

(i) $y \sim N_n(X\beta, \sigma^2 I)$.

(ii) $\hat{\beta} \sim N_p(\beta, \sigma^2(X^T X)^{-1})$.

- A média e a estrutura de covariância de $\hat{\beta}$ foram obtidas em (ii). A normalidade decorre do fato de ser uma função linear do vetor y, cuja distribuição é normal

Análise de Variância (ANOVA)

- Técnica mais usada para verificar a adequação do ajuste do modelo de regressão a um conjunto de dados.
- Baseia-se na seguinte identidade

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

Variabilidade Total (SQT) = Variabilidade Explicada (SQE) + Soma de Quadrados Residual (SQR)

ANOVA

- O *coeficiente de correlação múltipla de Pearson* (ou *coeficiente de determinação*) **R^2** expressa o quanto o modelo explica a variabilidade total da variável *y*.

$$R^2 = \frac{SQE}{SQT} = \frac{\sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\boxed{\textit{Variabilidade Explicada (SQE)}}}{\boxed{\textit{Variabilidade Total (SQT)}}}$$

ANOVA

$$R^2(ajustado) = \frac{SQE/(n-p)}{SQT/(n-1)}$$

ANOVA

- CUIDADO: Alguns pesquisadores se baseiam erroneamente **apenas** no valor de R^2 para escolher o melhor modelo. Tão importante quanto ter um R^2 próximo a 1, é que a estimativa de σ^2 também seja pequena, pois os intervalos de confiança para os parâmetros de interesse são proporcionais a σ.

Tabela ANOVA

A tabela da ANOVA é usada para testar a *adequação global* do modelo de regressão

$$y_{(nx1)} = X_{(nxp)}\beta_{(px1)} + \varepsilon_{(nx1)}$$

Efeito	Soma de Quadrados	GL	Média de Quadrados	Estatística
Regr.	SQE	p-1	MQE=SQE/(p-1)	F = MQE/MQR
Residual	SQR	n-p	MQR=SQR/(n-p)	
Total	SQT	n-1		

Teste F - *Adequação Global*

- Hipóteses:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1: $\beta_1 \neq 0$ ou $\beta_2 \neq 0$ ou ... ou $\beta_p \neq 0$. (pelo menos um)

- Estatística do Teste

$$\mathbf{F = MQE / MQR}$$

- Se $F > F_{p-1,\, n-p}(\alpha)$ rejeita H_0 => o efeito global de pelo menos algumas variáveis presentes na matriz X explica a variabilidade de *y*.

Teste F - *Adequação Global*

- A estatística do teste **F** representa o quociente entre **SQE** e **SQR** que têm distribuição χ^2, pelos respectivos *g.l.*
- Por isso, temos que

$$\mathbf{F \sim F_{p-1,\, n-p}(\alpha)},$$

que representa o valor de uma distribuição F-Snedecor com *p-1* e *n-p* graus de liberdade, ao nível de significância α

Seleção das Variáveis Explicativas – Teste t

- O **Teste F** permite apenas concluir que algumas variáveis explicativas são realmente importantes (mas não sabemos quais!! ☹).

- O **Teste t** permite selecionar as variáveis independentes (explicativas) que são significativas para o modelo.

Seleção das Variáveis Explicativas – Teste t

- Eliminar variáveis que tem pouca ou nenhuma contribuição na variabilidade da variável dependente *y*.

- Hipóteses:

$$H_0\text{: } \beta_r = 0$$

$$H_1\text{: } \beta_r \neq 0, \qquad r = 0, \ldots, p$$

Seleção das Variáveis Explicativas – Teste t

- Estatística do Teste

$$T_r = \frac{|\hat{\beta}_r|}{\hat{\sigma}\sqrt{v_{rr}}}$$

- Se $T_r > t_{n-p}(\alpha)$, rejeita H_0 => a variável independente x_r é significativa para explicar a variabilidade da resposta *y* e deve permanecer no modelo.

Seleção das Variáveis Explicativas – Teste t

- Note que

$$T_r = \frac{|\hat{\beta}_r|}{\hat{\sigma}\sqrt{v_{rr}}}$$

$\hat{\beta}_r \sim N(\beta_r, \sigma^2 v_{rr})$, onde v_{rr} é o elemento (r,r) da diagonal de $(X^T X)^{-1}$

$\hat{\sigma}^2 \sim (n-p)^{-1}\chi^2_{n-p}$

- Note que a estatística do teste T_r representa o quociente uma distribuição **Normal-Padrão** pela raiz quadrada de uma **distribuição** χ^2 pelo respectivo *g.l.*
- Por isso $T_r \sim t_{n-p}(\alpha)$, representa o valor de uma distribuição t-Student com *n-p* graus de liberdade, ao nível de significância α.

Exemplo

- Objetivo: estimar o consumo de combustível nos estados americanos;
- Variável dependente (*y*):
  - Cons = consumo de gasolina (c/gl)
- Variáveis independentes (X)
  - Taxa = valor do imposto estadual;
  - Rend = renda média em US$;
  - Rodov = extensão da malha rodoviária estadual;
  - Licen = % da população habilitada a dirigir.

Exemplo - BD

<u>Base de Dados</u>

Consumo de Combustível nos estados americanos

Con	Tax	Ren	Rod	Lic	Con	Tax	Ren	Rod	Lic
541	9.0	3571	1976	52.5	460	8.5	4574	2619	55.1
524	9.0	4092	1250	57.2	566	9.0	3721	4746	54.4
561	9.0	3865	1586	58.0	577	8.0	3448	5399	54.8
414	7.5	4870	2351	52.9	631	7.5	3846	9061	57.9
410	8.0	4399	431	54.4	574	8.0	4188	5975	56.3
457	10.0	5342	1333	57.1	534	9.0	3601	4650	49.3
344	8.0	5319	11868	45.1	571	7.0	3640	6905	51.8
467	8.0	5126	2138	55.3	554	7.0	3333	6594	51.3
464	8.0	4447	8577	52.9	577	8.0	3063	6524	57.8
498	7.0	4512	8507	55.2	628	7.5	3357	4121	54.7
580	8.0	4391	5939	53.0	487	8.0	3528	3495	48.7
471	7.5	5126	14186	52.5	644	6.5	3802	7834	62.9
525	7.0	4817	6930	57.4	640	5.0	4045	17782	56.6
508	7.0	4207	6580	54.5	704	7.0	3897	6385	58.6
566	7.0	4332	8159	60.8	648	8.5	3635	3274	66.3
635	7.0	4318	10340	58.6	968	7.0	4345	3905	67.2
603	7.0	4206	8508	57.2	587	7.0	4449	4639	62.6
714	7.0	3718	4725	54.0	699	7.0	3656	3985	56.3
865	7.0	4716	5915	72.4	632	7.0	4300	3635	60.3
640	8.5	4341	6010	67.7	591	7.0	3745	2611	50.8
649	7.0	4593	7834	66.3	782	6.0	5215	2302	67.2
540	8.0	4983	602	60.2	510	9.0	4476	3942	57.1
464	9.0	4897	2449	51.1	610	7.0	4296	4083	62.3
547	9.0	4258	4686	51.7	524	7.0	5002	9794	59.3

Exemplo: Ajuste (p1)

```
The regression equation is
Cons = 375 - 34.5 Taxa - 0.0665 Rend - 0.00240 Rodov + 13.4
Licen

Predictor        Coef       StDev          T        P
Constant        374.7       185.7       2.02    0.050
Taxa           -34.52       12.97      -2.66    0.011
Rend         -0.06653     0.01724      -3.86    0.000
Rodov       -0.002399    0.003394      -0.71    0.483
Licen          13.367       1.927       6.94    0.000

S = 66.38        R-Sq = 67.8%      R-Sq(adj) = 64.8%

Analysis of Variance

Source        DF          SS          MS         F        P
Regression     4      398906       99726     22.63    0.000
Error         43      189461        4406
Total         47      588366
```


Regra do p-value (p)

Se $p < \alpha$, rejeita H_0 ao nível de significância α%

Exemplo: Ajuste (p2)

```
The regression equation is
Cons = 305 - 29.3 Taxa - 0.0680 Rend + 13.7 Licen

Predictor        Coef       StDev          T        P
Constant        305.5       156.9       1.95    0.058
Taxa           -29.28       10.58      -2.77    0.008
Rend         -0.06796     0.01703      -3.99    0.000
Licen          13.747       1.839       7.47    0.000

S = 66.00        R-Sq = 67.4%     R-Sq(adj) = 65.2%
Analysis of Variance
Source        DF          SS          MS          F        P
Regression     3      396705      132235      30.36    0.000
Error         44      191662        4356
Total         47      588366
```

Regra do p-value (p)

Se $p < \alpha$, rejeita H_0 ao nível de significância α%

Exemplo: Ajuste (p3)

The regression equation is
Cons = - 15.2 Taxa - 0.0575 Rend + 16.4 Licen

Predictor	Coef	StDev	T	P
Noconstant				
Taxa	-15.172	7.939	-1.91	0.062
Rend	-0.05751	0.01665	-3.45	0.001
Licen	16.410	1.267	12.95	0.000

S = 68.01

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	16348097	5449366	1177.99	0.000
Error	45	208170	4626		
Total	48	16556267			

Regra do p-value (p)

Se $p < \alpha$, rejeita H_0 ao nível de significância α%

Exemplo: Ajuste (p4)

The regression equation is
Cons = - 0.0703 Rend + 15.3 Licen

Predictor	Coef	StDev	T	P
Noconstant				
Rend	-0.07035	0.01567	-4.49	0.000
Licen	15.344	1.170	13.11	0.000

S = 69.95

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	16331202	8165601	1668.93	0.000
Error	46	225065	4893		

Regra do p-value (p)

Se $p < \alpha$, rejeita H_0 ao nível de significância α%

Predição

- Um dos objetivos da análise de regressão
- Para um determinado vetor de valores

$$x_0^T = (1, x_{01}, x_{02}, \ldots, x_{0p})$$

de X, queremos prever o valor y_0 que deverá ser assumido pela variável resposta Y.

$$\hat{y}_0 = x_0^T \hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \ldots + \hat{\beta}_p x_{0p}$$

Predição

- Um intervalo com 100(1-α)% de confiança para o valor futuro y_0 é dado por

$$IC[(1-\alpha)\%, y_0] = \hat{y}_0 \pm t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 (1 + x_0^T (X^T X)^{-1} x_0)}$$