Regressão Linear

Ronald Targino Nojosa

DEMA-UFC

Notas de Aula - Parte 2
Versão Parcial

1. Regressão Linear Simples
   - Somas de Quadrados
   - Estimação de σ^2
   - Graus de Liberdade (gl)
   - Quadrados Médios
   - Tabela da ANOVA
   - Testes de Hipóteses para β_0 e β_1

Regressão Linear Simples

Somas de Quadrados

Pergunta: Quanto da variação nos dados é explicada pela reta de regressão? Em outras palavras, o quanto da variação de Y pode ser explicada pela variação de X?

Para responder, seguiremos a partir da identidade

$$Y_i - \hat{Y}_i = (Y_i - \overline{Y}) - (\hat{Y}_i - \overline{Y}).$$

Note que

$$(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i) \tag{10}$$

- $(Y_i - \overline{Y})$: distância entre o valor observado e a média geral
- $(Y_i - \hat{Y}_i)$: distância entre o valor observado e o predito pelo MRLS
- $(\hat{Y}_i - \overline{Y})$: distância entre o valor predito e a média geral[1]

[1] a média geral seria o valor predito caso os Y_i's fossem iid, o que implicaria a não necessidade de regressão

Segue:

$$(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i) \Rightarrow$$
$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \qquad (11)$$

As **somas de quadrados**[2] em (11) serão denotadas por

$$\boxed{SQ_{Tot} = SQ_{Reg} + SQ_{Res}}$$

[2] igualdade válida para modelos em que se verifica a equação (12); Tot: Total, Reg: Regressão, Res: Resíduo

Fazendo uso dos resultados vistos na seção Resíduos, o 3^o termo no desenvolvimento para a obtenção da equação (11) é nulo:

$$2\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i) = 2\left(\sum_{i=1}^{n}\hat{Y}_i e_i - \overline{Y}\sum_{i=1}^{n} e_i\right) = 0. \qquad (12)$$

A relação em (11) mostra que a variação dos valores de Y em torno da sua média ($\sum(Y_i - \overline{Y})^2$) pode ser dividida em duas partes: uma explicada pela regressão ($\sum(\hat{Y}_i - \overline{Y})^2$) e outra não explicada pela regressão ($\sum(Y_i - \hat{Y}_i)^2$), que se deve ao fato de que nem todas os pontos estão sobre a reta de regressão.

Identificação das **Somas de Quadrados (SQ)**[3]

$$SQ_{Tot} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = S_{YY}$$

$$SQ_{Reg} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \hat{\beta}_1^2 S_{XX} = \hat{\beta}_1 S_{XY} = S_{XY}^2/S_{XX}$$

$$SQ_{Res} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = S_{YY} - S_{XY}^2/S_{XX}$$

[3]Tot: Total, Reg: Regressão, Res: Resíduo

Propostas das **Somas de Quadrados (SQ)**

- SQ_{Tot}: medir a variação total dos valores observados de Y em relação a sua média.
- SQ_{Reg}: medir a parcela da variabilidade de Y que é explicada pelo modelo de regressão
- SQ_{Res}: medir a parcela da variabilidade de Y que não é explicada pelo modelo de regressão

Pergunta: Qual a proporção da variabilidade de Y que é explicada pelo MRLS?

Temos:

$$SQ_{Tot} = SQ_{Reg} + SQ_{Res} \Rightarrow$$
$$1 = \frac{SQ_{Reg}}{SQ_{Tot}} + \frac{SQ_{Res}}{SQ_{Tot}}$$

Sendo, $\frac{SQ_{Reg}}{SQ_{Tot}} \geq 0$ e $\frac{SQ_{Res}}{SQ_{Tot}} \geq 0$, concluímos que

$$0 \leq \frac{SQ_{Reg}}{SQ_{Tot}} \leq 1$$

Coeficiente de Determinação R^2

R^2 representa a proporção da variabilidade da variável resposta que é explicada pelo modelo de regressão. Por definição temos:

$$R^2 = \frac{SQ_{Reg}}{SQ_{Tot}}, \tag{13}$$

com $0 \leq R^2 \leq 1$. Um valor de R^2 próximo a 1 não garante um bom ajuste do modelo de regressão.

Exercícios P2 - 1.1.

1. *Mostre:*
   a. $\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$
   b. $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}[\hat{\beta}_1(X_i - \overline{X})]^2 = \frac{S_{XY}^2}{S_{XX}}$
   c. $R^2 = r^2$, *em que* r *é o coeficiente de correlação linear de Pearson.*

Estimação de σ^2

Na busca de um estimador para σ^2, vamos determinar as **Esperanças das Somas de Quadrados**:

Exercícios P2 - 1.2.

1. *Mostre que:*
   a. $E(SQ_{Tot}) = \beta_1^2 S_{XX} + (n-1)\sigma^2$
   b. $E(SQ_{Reg}) = E(\hat{\beta_1}^2 S_{XX}) = S_{XX} E(\hat{\beta_1}^2) = \sigma^2 + \beta_1^2 S_{XX}$
   c. $E(SQ_{Res}) = (n-2)\sigma^2$

Estimador de σ^2

Note que $E(SQ_{Res}) = (n-2)\sigma^2$. Assim, o estimador não viesado para σ^2 é

$$\hat{\sigma}^2 = \frac{SQ_{Res}}{n-2}. \tag{14}$$

Graus de Liberdade (gl)

Os **graus de liberdade** indicam quantos valores/termos são livres para variar após a "imposição de restrições" para a estimação das quantidades de interesse.

- Associados a $SQ_{Tot} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ temos $(n-1)$ gl.

  A SQ_{Tot} é uma função dos n termos $(Y_i - \overline{Y})$. Para sua obtenção, 1 grau de liberdade é perdido devido a restrição $\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$ sobre os desvios $(Y_i - \overline{Y})$.

- Associados a $SQ_{Reg} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ temos 1 gl.

  Mostramos que $SQ_{Reg} = \hat{\beta}_1^2 S_{XX}$, portanto apresenta 1 único grau de liberdade relacionado a obtenção de $\hat{\beta}_1$.

- Associados a $SQ_{Res} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ temos $(n-2)$ gl.

  A SQ_{Res} é uma função dos n termos $(Y_i - \hat{Y}_i)$. Duas restrições são impostas sobre os resíduos $(Y_i - \hat{Y}_i)$ como resultado da estimação de $\hat{\beta}_0$ e $\hat{\beta}_1$.

Propriedade da aditividade dos graus de liberdade

$$\boxed{(n - 1) = 1 + (n - 2)}$$

Quadrados Médios

Uma soma de quadrado dividida por seus graus de liberdade é denominada **Quadrado Médio (QM)**. De particular interesse na regressão, temos:

$$QM_{Reg} = \frac{SQ_{Reg}}{1} = SQ_{Reg}$$
$$QM_{Res} = \frac{SQ_{Res}}{n-2}$$

Observações:

- $\hat{\sigma}^2 = QM_{Res}$ Equação (14)
- QMs não são aditivos,isto é, $QM_{Tot} \neq QM_{Reg} + QM_{Res}$

Exercícios P2 - 1.3.

1. *Mostre que:*
   a. $E(QM_{Reg}) = \sigma^2 + \beta_1^2 S_{XX}$
   b. $E(QM_{Res}) = \sigma^2$

Teste F

Análise de Variância
Analysis of Variance - ANOVA

Os resultados vistos anteriormente são estruturados na denominada Tabela da Análise de Variância (Tabela ANOVA).

Tabela da Análise de Variância do MRLS

Fonte de variação	gl	SQ	QM	F_0
Regressão	1	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	QM_{Reg}	QM_{Reg}/QM_{Res}
Resíduo	$n-2$	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	QM_{Res}	
Total	$n-1$	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$		

Nota: Abordaremos F_0 posteriormente. Se o valor observado de QM_{Reg}/QM_{Res} for grande, temos indicação para $\beta_1 \neq 0$, caso contrário, a indicação é para $\beta_1 = 0$, isto é, não há influência linear de X sobre Y.

Testes de Hipóteses para β_0 e β_1

Para testar hipóteses e construir intervalos de confiança é necessário considerar/utilizar a **suposição** de que os erros seguem uma distribuição normal de probabilidade. Portanto,

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Exercícios P2 - 1.4.

1. *Para* $Y \sim N(\mu, \sigma^2)$, *com função geradora de momentos* $M_Y(t) = E(e^{tY}) = e^{t\mu + \frac{t^2}{2}\sigma^2}$, *e* $\{Y_1, \ldots, Y_n\}$ *uma aas de* Y, *determine:*
   a. $M_W(t)$, *em que* $W = a + bY$, *a e b constantes.*
   b. $M_S(t)$, *em que* $S = \sum_{i=1}^{n} Y_i$.
   c. $M_{\overline{Y}}(t)$, *em que* $\overline{Y} = S/n$.
2. *Para* $Y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, *independentes, determine* $M_S(t)$, *em que* $S = \sum_{i=1}^{n} Y_i$.
3. *Para* $Y_i \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, *independentes, determine* $M_S(t)$, *em que* $S = \sum_{i=1}^{n} a_i Y_i$, a_i*'s constantes.*

Um resultado importante para prosseguir com a parte inferencial é o **Teorema de Cochran**, que, sob as condições $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ e $\beta_1 = 0$, fornece condições necessárias e suficientes para estabelecer:

i. $\frac{SQ_{Tot}}{\sigma^2} \sim \chi^2_{(n-1)}$

ii. $\frac{SQ_{Reg}}{\sigma^2} \sim \chi^2_1$

iii. $\frac{SQ_{Res}}{\sigma^2} \sim \chi^2_{(n-2)}$

iv. SQ_{Res} e SQ_{Reg} são independentes.

Observando F_0 na tabela da ANOVA, vemos que:

$$\frac{QM_{Reg}}{QM_{Res}} = \frac{SQ_{Reg}/1}{SQ_{Res}/(n-2)} = \frac{\frac{SQ_{Reg}}{\sigma^2}/1}{\frac{SQ_{Res}}{\sigma^2}/(n-2)} = \frac{\chi^2_1/1}{\chi^2_{(n-2)}/(n-2)}.$$

Portanto, sob $\beta_1 = 0$ e normalidade dos erros,

$$F_0 = \frac{QM_{Reg}}{QM_{Res}} \sim F_{(1,\ n-2)}. \tag{15}$$

Note que os valores esperados para QM_{Reg} e QM_{Res} sinalizam que se o valor observado de F_0 for grande, então, provavelmente, $\beta_1 \neq 0$.

Para $\beta_1 \neq 0$, demonstra-se que

$$F_0 = \frac{QM_{Reg}}{QM_{Res}} \sim F_{(1,\ n-2,\ \lambda)},$$

isto é, F_0 tem distribuição F não central com parâmetro de não-centralidade $\lambda = \beta_1^2 S_{XX}/\sigma^2$.

Este desenvolvimento nos permite testar as hipóteses

$$H_0 : \beta_1 = 0 \quad \textit{versus} \quad H_1 : \beta_1 \neq 0$$

com base na estatística F_0, rejeitando H_0, sob um nível de significância $\alpha \in (0, 1)$, se [4]

$$F_0 > F_{(1-\alpha;\ 1,\ n-2)}.$$

O nível descritivo do teste (valor-p; *p value*) será dado por

$$P(F_{(1,\ n-2)} > F_0).$$

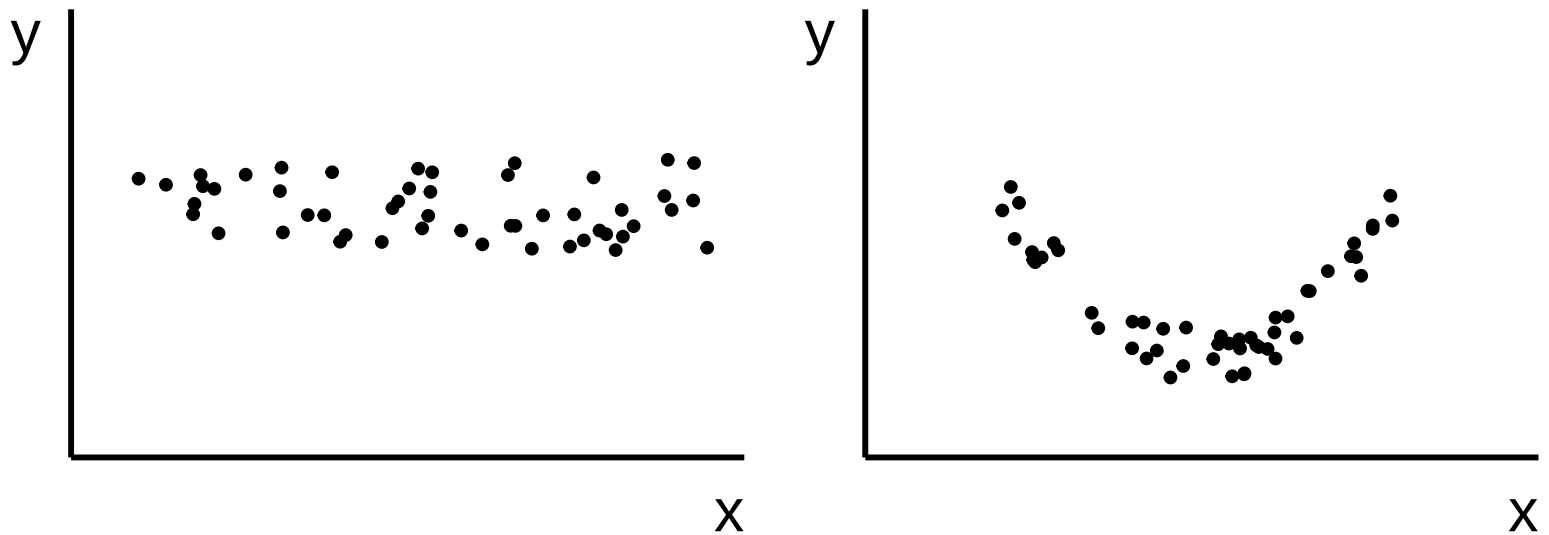
Este é o teste da **Significância da Regressão** para o MRLS.

[4] para H_0 falsa, $E(QM_{Reg}) > E(QM_{Res})$ E(QM)

Significância da regressão

Hipóteses: $H_0 : \beta_1 = 0 \quad \times \quad H_1 : \beta_1 \neq 0$

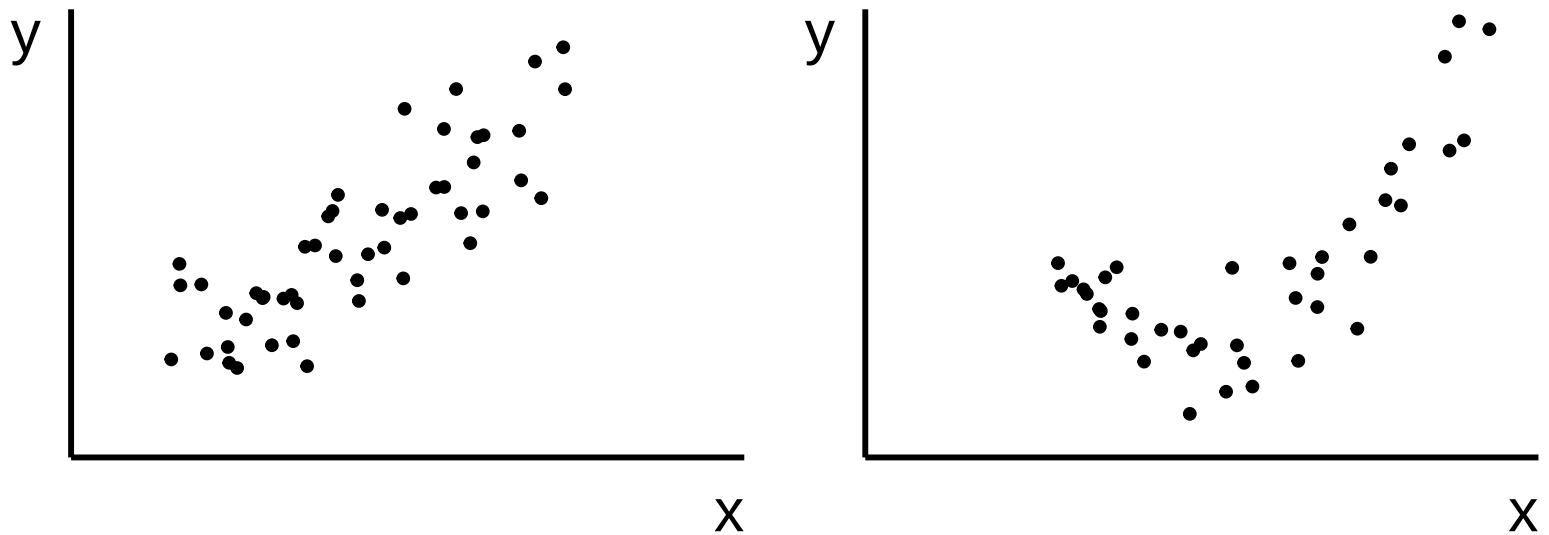
Não rejeitar $H_0 \Rightarrow$ não há relação linear entre X e Y



Significância da regressão

Hipóteses: $H_0 : \beta_1 = 0 \quad \times \quad H_1 : \beta_1 \neq 0$

Rejeitar $H_0 \Rightarrow$ há relação linear entre X e Y



Teste para o parâmetro β_1

Hipóteses: $H_0: \ \beta_1 = \beta_{1_0}$ *versus* $H_1: \ \beta_1 \begin{matrix} < \\ \neq \\ > \end{matrix} \beta_{1_0}$

Estimador: $\hat{\beta}_1 \sim N\left(\beta_{1_0}, \dfrac{\sigma^2}{S_{xx}}\right)$

Variável de teste: $T_0 = \dfrac{\hat{\beta}_1 - \beta_{1_0}}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}}$, com $\hat{\sigma}^2 = QM_{Res}$.

Sob H_0, $T_0 \sim t(n-2)$.

Teste para o parâmetro β_1

Regra de Decisão

Hipótese alterntiva	H_0 será rejeitada se
$H_1: \beta_1 \neq \beta_{1_0}$	$\lvert t_0 \rvert > t_{(1-\frac{\alpha}{2}),(n-2)}$
$H_1: \beta_1 < \beta_{1_0}$	$t_0 < t_{(\frac{\alpha}{2}),(n-2)}$
$H_1: \beta_1 > \beta_{1_0}$	$t_0 > t_{(1-\frac{\alpha}{2}),(n-2)}$

Nota: t_0 é o valor observado de T_0; $t(p,(n-2))$ é o quantil de ordem p, $0 < p < 1$, de T_0; α é o nível de significância do teste.

Exercícios P2 - 1.5.

1. *Vimos que* $E(\hat{\beta}_1) = \beta_1$ *e* $V(\hat{\beta}_1) = \sigma^2/S_{XX}$. *Mostre:*

   a. $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{XX})$

   b. $T_0 \sim t(n-2)^a$

2. *Para cada um dos 3 casos (hipóteses alternativas) indique como calcular o nível descritivo do teste.*

---

[a] $t(v)$ é a razão entre uma $N(0,1)$ e a raiz de uma $\frac{\chi^2_v}{v}$, independentes

O teste para $\beta_1 = 0$ feito com base na estatística T_0 é equivalente ao teste da significância da regressão, no caso do MRLS, baseado na estatística F_0. Teremos

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}.$$

Verifique a equivalência mostrando que $T_0^2 = F_0$.

Teste para o parâmetro β_0

Hipóteses: $H_0: \ \beta_0 = \beta_{0_0} \quad \textit{versus} \quad H_1: \ \beta_0 \overset{<}{\underset{>}{\neq}} \beta_{0_0}$

Estimador: $\hat{\beta}_0 \sim N\left(\beta_{0_0}, \dfrac{\sigma^2}{n} + \dfrac{\overline{X}^2\sigma^2}{S_{XX}}\right)$

Variável de teste: $T_0 = \dfrac{\hat{\beta}_0 - \beta_{0_0}}{\sqrt{\dfrac{\hat{\sigma}^2}{n} + \dfrac{\overline{X}^2\hat{\sigma}^2}{S_{XX}}}}$, com $\hat{\sigma}^2 = QM_{Res}$.

Sob H_0, $T_0 \sim t(n-2)$.

Teste para o parâmetro β_0

Regra de Decisão

Hipótese alterntiva	H_0 será rejeitada se
$H_1: \beta_0 \neq \beta_{0_0}$	$\lvert t_0 \rvert > t_{(1-\frac{\alpha}{2}),(n-2)}$
$H_1: \beta_0 < \beta_{0_0}$	$t_0 < t_{(\frac{\alpha}{2}),(n-2)}$
$H_1: \beta_0 > \beta_{0_0}$	$t_0 > t_{(1-\frac{\alpha}{2}),(n-2)}$

Nota: t_0 é o valor observado de T_0; $t(p,(n-2))$ é o quantil de ordem p, $0 < p < 1$, de T_0; α é o nível de significância do teste.

Exercícios P2 - 1.6.

1. *Vimos que* $E(\hat{\beta}_0) = \beta_0$ *e* $V(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\overline{X}^2\sigma^2}{S_{xx}}$. *Mostre:*
   a. $\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{n} + \frac{\overline{X}^2\sigma^2}{S_{xx}})$
   b. $T_0 \sim t(n-2)^a$
2. *Para cada um dos 3 casos (hipóteses alternativas) indique como calcular o nível descritivo do teste.*

---

[a] $t(v)$ é a razão entre uma $N(0,1)$ e a raiz de uma $\frac{\chi^2_v}{v}$, independentes