

## Intervalos de confiança

Vimos que o estimador de verossimilhança para uma variável de Bernoulli é dado por

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad X_i \in \{0, 1\}$$

e  $P(X_i) = p^{X_i} (1-p)^{1-X_i}$ . Intervalos de confiança são úteis para estimar o que próximo estamos no verdadeiro valor  $p$ . Observe que logicamente isso parece estranho já que não conhecemos, a princípio, o valor de  $p$ . De maneira mais precisa, o que desejuríamos é a probabilidade do valor de  $p$  estar num certo intervalo. Entretanto, esse tipo de pergunta não tem sentido dentro da nossa forma de pensar. Iremos perguntar, portanto, algo como: se o experimento for realizado muitas vezes, qual a probabilidade de  $\hat{p}$  estar dentro de um certo intervalo.

Uma maneira para encontrar esse resultado é usando a desigualdade de Hoeffding:

$$P(|\hat{P}_n - p| > \epsilon) \leq 2 e^{-2n\epsilon^2}$$

Nesse caso, o intervalo de confiança ficaria da forma

$$I = [\hat{P}_n - \epsilon_n, \hat{P}_n + \epsilon_n]$$

sendo  $\epsilon_n$  escolhido de modo que

$$2 e^{-2n\epsilon^2} = \alpha$$

$$-2n\epsilon^2 = \ln(\alpha/2)$$

$$\epsilon^2 = \frac{1}{2n} \ln(\alpha/2)$$

$$\epsilon = \sqrt{\frac{1}{2n} \ln\left(\frac{\alpha}{2}\right)}$$

e, portanto,

$$P(\rho \notin I) = P(|\hat{P}_n - \rho| > \epsilon_n) \leq \alpha$$

ou, de outro modo

$$P(\rho \in I) = P(|\hat{P}_n - \rho| < \epsilon_n) = 1 - P(|\hat{P}_n - \rho| > \epsilon_n) \geq 1 - \alpha$$

Tomando um exemplo numérico para  $n=100$  e  $\alpha=0.05$ , temos

$$\epsilon_n = \sqrt{\frac{1}{2 \times 100} \ln\left(\frac{2}{0.05}\right)} \approx 0,13581$$

Assim, esperamos que em 95% das vezes, o valor de  $\hat{P}$  esteja no intervalo

$$[\hat{P} - 0,136, \hat{P} + 0,136].$$

### Exemplo no notebook

Porém, a maneira mais "usual" para estimar esses intervalos de confiança é usar a hipótese de normalidade assintótica.

Nesse caso, temos o erro padrão

$$se = \sqrt{V(\hat{\theta}_n)}$$

sendo  $\hat{\theta}_n$  o estimador para  $\theta$ , calculado para uma amostra de tamanho  $n$   $\{x_1, x_2, \dots, x_n\}$  e  $V(\hat{\theta}_n)$  a variância de  $\hat{\theta}_n$ . No caso da variável de Bernoulli, já vimos que

$$\hat{p} = \frac{1}{n} \sum x_i \quad e \quad V(\hat{p}_n) = p(1-p)/n$$

Logo, usando o princípio plugging, temos

$$\hat{se} = \sqrt{\hat{p}(1-\hat{p})/n}$$

Além disso, usando o fato que estimadores de

verossimilhancer são assintoticamente normais, temos que

$$\hat{P}_n \sim N(p, \hat{s}^2 e^2)$$

Se estamos interessados no intervalo de confiança  $(1-\alpha)$ , temos que encontramos  $\xi$ , tal que,

$$P(|\hat{P}_n - p| < \xi) \Rightarrow 1 - \alpha$$

Visto que  $\hat{P}_n \sim N(p, \hat{s}^2 e^2)$ , temos que

$$X = \hat{P}_n - p \sim N(0, \hat{s}^2 e^2)$$

Dessa modo, devemos resolver

$$\int_{-\xi}^{\xi} N(0, \hat{s}^2 e^2) dx = 1 - \alpha$$

$$Erf(x) = \frac{x}{\sqrt{2}} \int_0^x e^{-t^2} dt$$

$$Erf\left[\frac{\xi}{\sqrt{2} \hat{s} e}\right] = 1 - \alpha$$

$$\xi = \sqrt{2} \hat{s} e Erf^{-1}[1 - \alpha]$$

Uma vez encontrado  $\xi$ , sabemos que ~~é~~ a probabilidade de  $\hat{P}$  estar no intervalo  $[p - \xi, p + \xi]$  é  $(1 - \alpha)$ .

### Exemplo notebook

Intervalos de confiança e testes de hipótese

Intervalos de confiança e testes de hipótese  
Ocorre que existe uma relação clara entre testes de hipótese e intervalos de confiança. Para ver essa conexão, considere

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$$

Vimos que a região de rejeição de  $H_0$  pode ser escrita

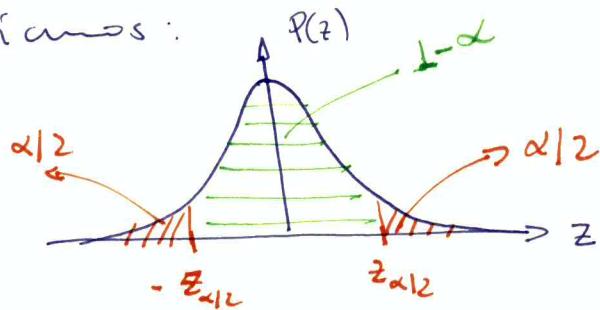
$$|\bar{X} - \mu_0| > Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

sendo

$$P(z > z_{\alpha/2}) = \alpha/2$$

e  $P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha$  com  $z \sim N(0, 1)$ . Numa

gráfico, teríamos:



Note que essa condição para  $z = |\bar{x} - \mu_0|$  é equivalente a

$$\mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{x} < z_{\alpha/2} \frac{\sigma}{\sqrt{n}} + \mu_0$$

ou ainda

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Se o tamanho do teste é  $\alpha$ , a probabilidade do falso alarme (FPR - false positive rate)

$$P(\text{rejeitar } H_0 \mid \mu = \mu_0) = \alpha$$

Por outro lado, o TPR (true positive rate) é

$$P(\text{aceitar } H_0 \mid \mu = \mu_0) = 1 - \alpha$$

Usando a condição passada no intervalo de  $\mu_0$ , temos

$$P\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \mid \mu = \mu_0\right) = 1 - \alpha$$

Viu-se que a condição vale para qualquer  $\mu_0$ , podemos escrever

$$P\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Note que a expressão anterior define o intervalo de confiança de  $1-\alpha$  para  $\bar{X}$ . Assim, invertendo o intervalo de aceitação de  $H_0$ , encontramos o intervalo de confiança para  $\bar{X}$ . Vemos, portanto, que o teste da hipótese fixa o parâmetro ( $\mu = \mu_0$ ) e pergunta quais valores de  $\bar{X}$  são compatíveis com  $H_0$ . Por outro lado, o intervalo de confiança fixa os valor de  $\bar{X}$  e pergunta/define uma região mais provável para  $\bar{X}$ .

### Régressão linear

Podemos pensar numa regressão linear como uma relação entre o valor esperado condicional de uma variável  $Y$  dado que observamos  $X$ , isto é,

$$y = E(Y|X=x) \approx b + ax$$

sendo  $a$  e  $b$  os coeficientes do modelo linear,  $y$  a variável resposta (ou dependente) e  $x$  a variável independente ou regressora. Entretanto, como os valores de  $y$  são variáveis aleatórias é comum incluir um termo de ruído aditivo:

$$y = E(Y|X=x) + \epsilon_i \approx a x + b + \epsilon_i$$

sendo  $E(\epsilon_i) = 0$  e os  $\epsilon_i$  independentes e identicamente distribuídos.

Dado um conjunto de valores  $(x_i, y_i) \forall i \in \{1, 2, \dots, n\}$  nossa primeira tarefa é determinar a e b. Para isso, vamos assumir ainda que  $\epsilon_i \sim N(0, \sigma^2)$ . Sendo assim,

$$E(y) = E[a x + b + \epsilon_i]$$

$$= a x + b$$

Akém disso,

$$V(y) = V[\alpha x + b + \epsilon_i] \\ = \sigma^2$$

Vamos assumir que dado um  $x_i$ ,  $y_i$  é uma variável aleatória gerada por  $\epsilon_i$  com as características anteriores. Desse modo, podemos usar o procedimento de máxima verossimilhança, ou seja,

$$\ell(a, b) = \prod_{i=1}^n N(\alpha x_i + b, \sigma^2) \\ = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y_i - (\alpha x_i + b))^2}{2\sigma^2} \right\}$$

$$\ell(a, b) = \log \ell(a, b) \\ = \sum_{i=1}^n \ln \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \sum_{i=1}^n \frac{(y_i - \alpha x_i - b)^2}{2\sigma^2}$$

$$\ell(a, b) \propto -\sum_{i=1}^n (y_i - \alpha x_i - b)^2$$

Derivando com relação a  $a$  e  $b$ , temos

$$\frac{\partial \ell}{\partial a} = 2 \sum_{i=1}^n x_i (y_i - \alpha x_i - b) = 0$$

$$\Rightarrow \sum_{i=1}^n x_i (b + \alpha x_i - y_i) = 0$$

$$\frac{\partial \ell}{\partial b} = 2 \sum_{i=1}^n (y_i - \alpha x_i - b) = 0$$

$$\Rightarrow \sum_{i=1}^n (b + \alpha x_i - y_i) = 0$$

Ficamos com um sistema de equações

$$\sum_{i=1}^n x_i y_i = b \sum_{i=1}^n x_i + a \sum_{i=1}^n x_i^2$$

$$\sum_{i=1}^n y_i = n b + a \sum_{i=1}^n x_i$$

Depois que alguma álgebra:

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{b} = \bar{y} - \hat{a} \bar{x}$$

sendo

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad e \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

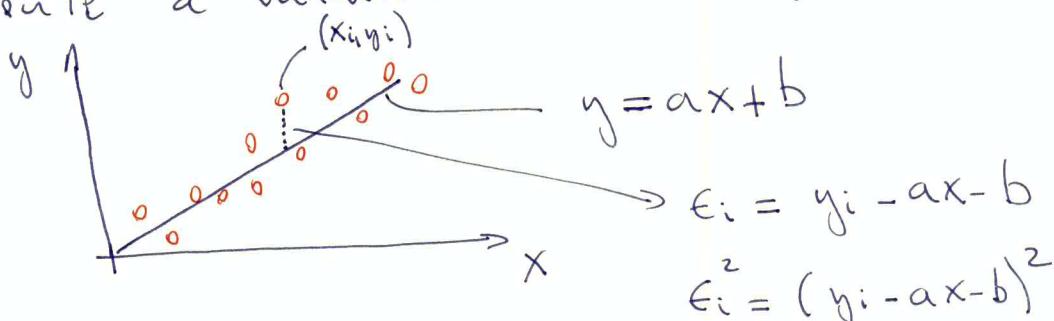
Poderíamos estimar ainda o  $\sigma$ , nesse caso,

$$\frac{\partial L}{\partial \sigma} = \frac{\partial}{\partial \sigma} \left\{ \sum_{i=1}^n \ln \sigma^{-1} - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - ax_i - b)^2 \right\}$$

$$= -\frac{1}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (y_i - ax_i - b)^2 = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - ax_i - b)^2$$

Vale notar que  $\sigma$  não compõe a parte determinística do modelo linear. Observe ainda que esse procedimento é equivalente a minimizar o erro quadrático médio.



Uma maneira alternativa para obter  $\hat{a}$  e  $\hat{b}$  é supor um conjunto de  $m$  valores de  $y_{i,k}$   $\{y_{i,k}\}_{k=1}^m$  e calcular

$$\hat{y}_i = \frac{1}{m} \sum_{k=1}^m y_{i,k}$$

Sabemos que  $E(y_i) = ax_i + b$ . Tomando uma outra amostra, temos

$$\hat{\bar{y}}_j = \frac{1}{m} \sum_{k=1}^m y_{j,k}$$

sendo  $E(y_j) = ax_j + b$ . Vamos ver que as estimativas para os valores esperados são

$$E(y_i) = \hat{y}_i$$

$$E(y_j) = \hat{\bar{y}}_j$$

Temos um sistema

$$\hat{y}_i = ax_i + b$$

$$\hat{\bar{y}}_j = ax_j + b$$

cujas soluções é

$$\hat{a} = \frac{\hat{y}_i - \hat{\bar{y}}_j}{x_i - x_j}$$

$$\hat{b} = \frac{x_i \hat{y}_j - x_j \hat{y}_i}{x_i - x_j}$$

Podemos calcular também os valores esperados de  $\hat{a}$  e  $\hat{b}$

$$E(\hat{a}) = \frac{ax_i + b - ax_j - b}{x_i - x_j} = a$$

$$E(\hat{b}) = \frac{x_i(a x_j + b) - x_j(a x_i + b)}{x_i - x_j}$$

$$= \frac{(x_i - x_j)b}{x_i - x_j} = b$$

Assim como suas variâncias (para  $x_j = 0$ )

$$V(\hat{a}) = \frac{2\sigma^2}{x_i^2}$$

$$V(\hat{b}) = \sigma^2$$

Notamos que os estimadores não apresentam bias e que a variância de  $\hat{a}$  diminui com  $x_i$ .

### Exemplo notebook

#### Regressão usando métodos de projeto

Usando notações vetoriais  $\bar{y} = (y_1, y_2, \dots, y_n)$  e  $\bar{x} = (x_1, x_2, \dots, x_n)$ , nosso modelo linear fica:

$$\bar{y} = a \bar{x} + b \bar{I} + \bar{\epsilon}$$

sendo  $\bar{I} = (1, 1, \dots, 1)$  e  $\bar{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)$ . Lembrando que

$$\langle \bar{x}, \bar{y} \rangle = \bar{E}(\bar{x}^T \bar{y})$$

Podemos tomar o produto escalar com um  $\bar{x}_1 \in \bar{I}^\perp$ , sendo  $\bar{I}^\perp$  um subespaço do espaço vetorial tal que um vetor desse subespaço segue  $\langle \bar{z}, \bar{I} \rangle = 0$ . Fazendo isso, temos

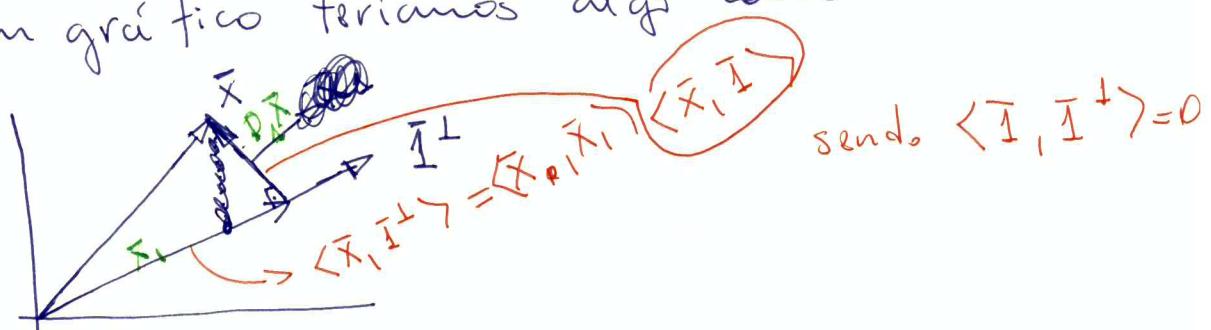
$$\langle \bar{y}, \bar{x}_1 \rangle = a \langle \bar{x}, \bar{x}_1 \rangle + b \cancel{\langle \bar{I}, \bar{x}_1 \rangle} \quad \cancel{\text{0}},$$

$$+ \langle \bar{\epsilon}, \bar{x}_1 \rangle$$

$$\hookrightarrow E(\bar{\epsilon}^T \bar{x}_1) = E(\bar{\epsilon}^T) \bar{x}_1 = 0 \quad (9)$$

$$\Rightarrow \hat{a} = \frac{\langle \bar{y}, \bar{x}_1 \rangle}{\langle \bar{x}, \bar{x}_1 \rangle}$$

Analisando o denominador, percebemos que ao projetar  $\bar{x}$  em  $\bar{I}^\perp$ , temos o minimal mean Squared error (MMSE), ou seja, num gráfico teríamos algo como



Desse modo, podemos pensar

$$\bar{x}_1 = P_{\bar{I}^\perp}(\bar{x})$$

Sendo  $P_{\bar{I}^\perp}$  o operador/matriz de projeção. Sendo assim, o módulo de  $\bar{x}_1$  é no máximo igual a  $\bar{x}$ , além disso, usando Pitágoras temos

$$\langle \bar{x}, \bar{x} \rangle^2 = \langle \bar{x}_1, \bar{I} \rangle^2 + \langle \bar{x}, \bar{x}_1 \rangle^2$$

$$\langle \bar{x}, \bar{x}_1 \rangle^2 = \langle \bar{x}, \bar{x} \rangle^2 - \langle \bar{x}, \bar{I} \rangle^2$$

O primeiro termo à esquerda é módulo quadrado de  $\bar{x}$ , direita já o último é o módulo quadrado de  $\bar{x}$  projetado ao longo de  $\bar{I}$ . Note que essa escolha de  $\bar{x}_1$ , tem o efeito de reduzir a variância de  $\hat{a}$ , visto que quanto mais alinhado com  $\bar{I}$  for  $\bar{x}$ , maior deve ser a variância de  $\hat{a}$ . Podemos pensar que quanto mais próximo de  $\bar{I}$ , mais constante é  $\bar{x}$ , e da nossa experiência com o caso 1d, sabemos que

o tamanho de  $\bar{x}$  tem o papel de diminuir a variância.

Analisando o numerador de  $\hat{a}$  e considerando que

~~que P\_I é o projeção ortogonal~~

$$\bar{x}_1 + P_{\bar{I}} \bar{x} = \bar{x}$$

$$\bar{x}_1 = \bar{x} - P_{\bar{I}} \bar{x}$$

temos

$$\langle \bar{y}, \bar{x}_1 \rangle = \langle \bar{y}, \bar{x} \rangle - \langle \bar{y}, P_{\bar{I}} \bar{x} \rangle$$

Definindo o operador  $P_{\bar{I}}$

$$P_{\bar{I}} = \frac{\bar{I} \bar{I}^T}{\|\bar{I}\|^2} = \frac{\bar{I} \bar{I}^T}{n}$$

$$\begin{aligned}\|\bar{I}\|^2 &= (\bar{I}, \bar{I}) (\bar{I}, \bar{I}) \\ &= 1^2 + 1^2 + \dots \\ &= n\end{aligned}$$

Podemos calcular

$$\begin{aligned}\langle \bar{y}, P_{\bar{I}} \bar{x} \rangle &= \bar{y}^T P_{\bar{I}} \bar{x} = \frac{\bar{y}^T \bar{I} \bar{I}^T \bar{x}}{n} \\ &= (\bar{y}^T \bar{I})(\bar{I}^T \bar{x}) / n \\ &= \left( \sum_{i=1}^n y_i \right) \left( \sum_{i=1}^n x_i \right) / n\end{aligned}$$

similarmente

$$\begin{aligned}\langle \bar{x}, P_{\bar{I}} \bar{x} \rangle &= \bar{x}^T P_{\bar{I}} \bar{x} = \bar{x}^T \bar{I} \bar{I}^T \bar{x} / n \\ &= (\bar{x}^T \bar{I})(\bar{I}^T \bar{x}) / n \\ &= \left( \sum_{i=1}^n x_i \right)^2 / n\end{aligned}$$

Desse modo ficamos com

$$\hat{a} = \frac{\langle \bar{y}, \bar{x}_1 \rangle}{\langle \bar{x}, \bar{x}_1 \rangle} = \frac{\langle \bar{y}, \bar{x} \rangle - \langle \bar{y}, P_{\bar{x}} \bar{x} \rangle}{\langle \bar{x}, \bar{x} \rangle - \langle \bar{x}, P_{\bar{x}} \bar{x} \rangle}$$
$$\hookrightarrow \bar{x}_1 = \bar{x} - P_{\bar{x}} \bar{x}$$

$$\hat{a} = \frac{\bar{y}^T \bar{x} - \left( \sum_{i=1}^n y_i \right) \left( \sum_{i=1}^n x_i \right) / n}{\bar{x}^T \bar{x} - \left( \sum_{i=1}^n x_i \right)^2 / n}$$

Por um procedimento similar, tomando o produto escalar com  $\bar{x}^\perp$ , tal que  $\langle \bar{x}, \bar{x}^\perp \rangle = 0$ , temos

$$\langle \bar{y}, \bar{x}^\perp \rangle = b \langle \bar{I}, \bar{x}^\perp \rangle$$

$$\hat{b} = \frac{\langle \bar{y}, \bar{x}^\perp \rangle}{\langle \bar{I}, \bar{x}^\perp \rangle}$$

$$\hat{b} = \frac{\bar{x}^T \left( \sum_{i=1}^n y_i \right) / n - \bar{x}^T \bar{x} \left( \sum_{i=1}^n x_i \right) / n}{\bar{x}^T \bar{x} - \left( \sum_{i=1}^n x_i \right)^2 / n}$$

Poderíamos calcular também as variâncias

$$V(\hat{a}) = \frac{\sigma^2}{\|\bar{x}\|^2 - n(\bar{x})^2}$$

$$V(\hat{b}) = \frac{\sigma^2}{n - \frac{(n\bar{x})^2}{\|\bar{x}\|^2}}$$

As expressões anteriores dependem de  $\sigma$  que estimamos anteriormente como

$$\hat{\sigma} = \sqrt{\sum (y_i - \hat{a}x_i - \hat{b})^2 / n}$$

Entretanto é mais comum usarmos

$$\hat{F}^2 = \frac{\text{RSS}}{n-2}$$

com

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{a}x_i - \hat{b})^2$$

sendo essa última quantidade a soma dos resíduos ao quadrado (residual sum of squares). A quantidade  $n-2$  representa o número de graus de liberdade, isto é, o tamanho da amostra menos o número de parâmetros no modelo. Sob a hipótese de ruído gaussiano a quantidade

$$\frac{\text{RSS}}{\sigma^2} \sim \text{Chi}(n-2).$$

Uma outra quantidade que pode ser definida é

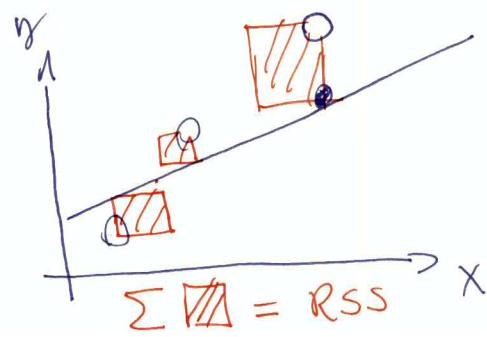
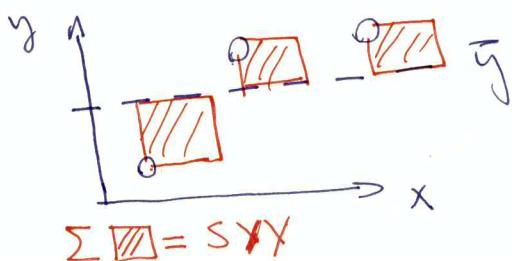
$$SYY = \sum_{i=1}^n (y_i - \bar{y})^2$$

a sum of squares about the mean. Combinando as duas anteriores, temos

$$R^2 = 1 - \frac{\text{RSS}}{SYY}$$

Essa quantidade é equivalente ao coeficiente de correlação de Pearson.

o chamado coeficiente de determinação. Num gráfico faríamos algo assim



Note que quanto melhor for o ajuste, mais próximo de  $R=1$  devemos estar. Por outro lado, se  $R=0$  devemos ter  $RSS = SYY$ , ou seja, uma reta horizontal, indicando que  $x$  não explica nada sobre  $y$ .

### Teste F

O teste F testa a diferença entre incluir o coeficiente a ou não. De modo mais específico:

$$H_0: E(Y|X=x) = b$$

$$H_1: E(Y|X=x) = b + \alpha x$$

A estatística do teste é

$$F = \frac{SYY - RSS}{\sum x^2}$$

Sob a hipótese de normalidade,  $F$  segue a distribuição F de Fisher-Snedecor

$$F(x; d_1, d_2) = \frac{\Gamma\left(\frac{d_1+d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)} \frac{d_1^{d_1/2} d_2^{d_2/2}}{(d_2 + d_1 x)^{\frac{d_1+d_2}{2}}} X$$

com  $d_1=1$  e  $d_2=n-2$ , isto é,

$$F \rightsquigarrow F(1, n-2)$$

Normalmente calculamos o p-valor associado a a F ser maior do que o valor empírico supondo  $H_0$  verdadeira.

E' comum definir ainda o  $R^2$  ajustado

$$\text{Adj. } R^2 = 1 - \frac{\text{RSS}/(n-p)}{\text{SYY}/(n-1)}$$

sendo  $n$  o tamanho da amostra e  $p$  o número de parâmetros no modelo. No caso do modelo linear  $p=2$ .

### Previsão linear

Supondo que um modelo linear foi ajustada a um conjunto de dados, produzindo

$$E(Y|X=x) \hat{=} \hat{a}x + \hat{b}$$

Se um novo valor de  $x$  for observado, digamos  $x_p$ , podemos prever o valor de  $y$  via

$$\hat{y}_p = \hat{a}x_p + \hat{b}$$

Podemos calcular também a variância

$$\begin{aligned} V(y_p) &= x_p^2 V(\hat{a}) + V(\hat{b}) + 2x_p V(\hat{a}\hat{b}) \\ &= x_p^2 V(\hat{a}) + V(\hat{b}) + 2x_p \text{Cov}(\hat{a}\hat{b}) \end{aligned}$$

O último termo é a covariância de  $\hat{a}$  e  $\hat{b}$ , e usando procedimentos similares aos do cálculo de  $V(\hat{a})$  e  $V(\hat{b})$  podemos mostrar que

$$V(\hat{a}\hat{b}) = \text{Cov}(\hat{a}\hat{b}) = \frac{-\bar{x}^2}{\sum_{i=1}^n (x_i^2 - \bar{x}^2)}$$

combinando esse resultado com as expressões para  $V(\hat{a})$  e  $V(\hat{b})$

térmos

$$V(y_p) = \hat{\sigma}^2 \frac{x_p^2 - 2x_p\bar{x} + \|\bar{x}\|^2/n}{\|\bar{x}\|^2 - n\bar{x}^2}$$

Vale notar ainda que devemos adicionar  $\hat{\sigma}^2$  a essa variância, uma vez que no valor de  $\hat{y}_p$  temos o termo do ruído extra  $\epsilon$ . Desse modo

$$\hat{\sigma}^2 = V(y_p) + \hat{\sigma}^2$$

e, consequentemente, o intervalo de confiança a 95% para  $y_p$  fica

$$P(-2 < N(0,1) < 2) \approx 0,95$$

$$P(y_p - 2\hat{\sigma} < y_p < y_p + 2\hat{\sigma}) \approx 0,95$$

Exemplo notebook

Extensão para múltiplas covariáveis

No caso em que temos múltiplas variáveis regressoras,

isto é,

$$y_i = x_1^{(i)}\alpha_1 + x_2^{(i)}\alpha_2 + \dots + x_p^{(i)}\alpha_p + \alpha_0 + \epsilon_i$$

podemos escrever

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

onde  $\bar{Y} = (y_1, y_2, \dots, y_n)^T$ ,  $\bar{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ ,  $\bar{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$

e

$$\bar{X} = \begin{pmatrix} \bar{x}_1^T \\ \bar{x}_2^T \\ \vdots \\ \bar{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_p^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_p^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \dots & x_p^{(n)} \end{pmatrix}$$

Nesse caso, poderíamos mostrar que o estimador para  $\bar{\beta}$  é

$$\hat{\bar{\beta}} = (\bar{X}^T \bar{X})^{-1} \bar{X}^T \bar{Y}$$

com variância

$$V(\hat{\bar{\beta}}) = \sigma^2 (\bar{X}^T \bar{X})^{-1}$$

Sub a hipótese de normalidade dos ruidos, temos

$$\hat{\bar{\beta}} \sim N(\bar{\beta}, \sigma^2 (\bar{X}^T \bar{X})^{-1})$$

sendo

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum \hat{\epsilon}_i^2$$

com  $\hat{\epsilon} = \bar{X} \bar{\beta} - \bar{Y}$ . Nesse contexto, costuma-se definir ainda a matriz de influência ou hat matrix como

$$\bar{V} = \bar{X} (\bar{X}^T \bar{X})^{-1} \bar{X}^T$$

representam o efeito de  $y_i$  para prever  $y_i$ ,  
pois

Note que  $\hat{Y} = \bar{V} \bar{Y}$

$$y_i = v_{1i} y_1 + \dots + v_{ii} y_i + \dots + v_{ni} y_n$$

Os elementos da diagonal de  $\bar{V}$  são chamados de leverage values e estão contidos no intervalo  $[0, 1]$ .

Esses valores medem a distância entre os  $x_i$  e seus valores médios em  $n$  observações. Podemos obter também a variância de cada resíduo como

$$V(\hat{y}_i - y_i) = V(\epsilon_i) = \sigma^2 (1 - v_{ii})$$

sendo  $\bar{v}_i$  os elementos da diagonal de  $\bar{Y}$ . Note que  $\bar{v}_i$  é no máximo 1, logo, a variância de  $e_i$  é sempre menor ou igual a  $\sigma^2$ .

Existe um problema de degenerescência quando duas ou mais colunas de  $\bar{X}$  são colineares. Uma possível solução para esse problema é tomar

$$\hat{\bar{B}} = (\bar{X}^T \bar{X} + \alpha I)^{-1} \bar{X}^T \bar{Y}$$

sendo  $\alpha > 0$  um parâmetro. Esse procedimento é conhecido como ridge regression e é equivalente a minimizar

$$\|\bar{Y} - \bar{X}\bar{B}\|^2 + \alpha \|\bar{B}\|^2$$

### Interpretando os resíduos

Nossa modelo linear assume um ruído aditivo normal. Essa hipótese pode ser verificada examinando os resíduos do ajuste

$$\hat{e}_i = \hat{a}x_i + \hat{b} - y_i$$

Uma possibilidade é usar o quantile-quantile plot. Podemos também analisar os resíduos normalizados

$$r_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - v_i}}$$

Essa quantidade deve estar distribuída como  $N(0,1)$ , de modo que a existência de valores  $r_i \notin [-1,96, 1,96]$  não deve exceder 5% dos dados, caso contrário a hipótese de normalidade e homocedasticidade (igualdade das variâncias) deve ser questionada.

O teste de Levene testa a hipótese da homoscedasticidade analisando a relação entre  $r_i$  e  $x_i$ , a qual não deve existir ( $r_i$  deve ser independente de  $x_i$ ).

### Rescala de variáveis

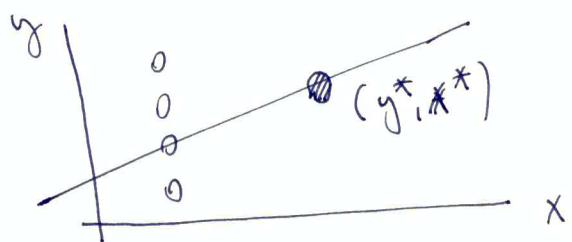
Após ajustar o modelo linear é tentador supor que valores de  $\bar{B}$  pequenos indicam variáveis de pouca importância. Entretanto, pode estar ocorrendo um problema de escala. Para contornar isso, um procedimento comum é normalizar as variáveis regressoras:

$$x'_i = \frac{x_i - \bar{x}}{\sigma_x}$$

Desse modo, devemos interpretar os valores de  $\bar{B}$  como coeficientes de correlação (isto é, limitados no intervalo  $-1$  a  $1$ ).

### Dados influentes

considere a figura abaixo



como está claro o ponto  $(x^*, y^*)$  é quem define a inclinação da reta, sendo, portanto muito influente na regressão linear em questão. A distância de cook é uma boa medida para quantificar a influência de um ponto na regressão linear.

Essa medida é definida como

$$D_i = \frac{\sum_j (y_{ji} - \hat{y}_{j(i)})^2}{(p/n) \sum_j (\hat{y}_{ji} - \bar{y}_j)^2}$$

sendo  $\hat{y}_{j(i)}$  o estimador de  $y$  ao não considerar o ponto  $(x_i, y_i)$  na regressão.

$D_i$  é maior que 1, devemos suspeitar que o ponto é um outlier, ainda que seja bem ajustado pelo modelo linear.

De maneira geral, se

Exemplo notebook