ME613 - Análise de Regressão

Parte 2

Benilton S Carvalho - 1S2020

Propriedades dos estimadores

Suposições do modelo de regressão linear simples

Até o momento, apenas suposições sobre esperança, variância e correlação foram feitas.

Desta forma, sabemos que os estimadores são não viesados e sabemos também quão precisos eles são.

No entanto, para construirmos intervalos e confiança, precisamos conhecer a distribuição de probabilidade desses estimadores.

Suposições do modelo de regressão linear simples

A partir de agora iremos assumir:

1. $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, 2, \ldots, n$
2. ε_i uma v.a. que segue a distribuição **Normal** em que $E(\varepsilon_i) = 0$ e $Var(\varepsilon_i) = \sigma^2$ desconhecida, para $i = 1, 2, \ldots, n$.
3. ε_i e ε_j são não-correlacionados para $i \neq j$, portanto $Cov(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$.

A suposição 3 implica em independência entre as observações i e j, $i \neq j$, no caso de normalidade. Desta maneira, utilizando as suposições 2 e 3, temos que ε_i's são iid.

Propriedades de Y_i

Já tínhamos visto que valor esperado para a resposta Y_i é:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

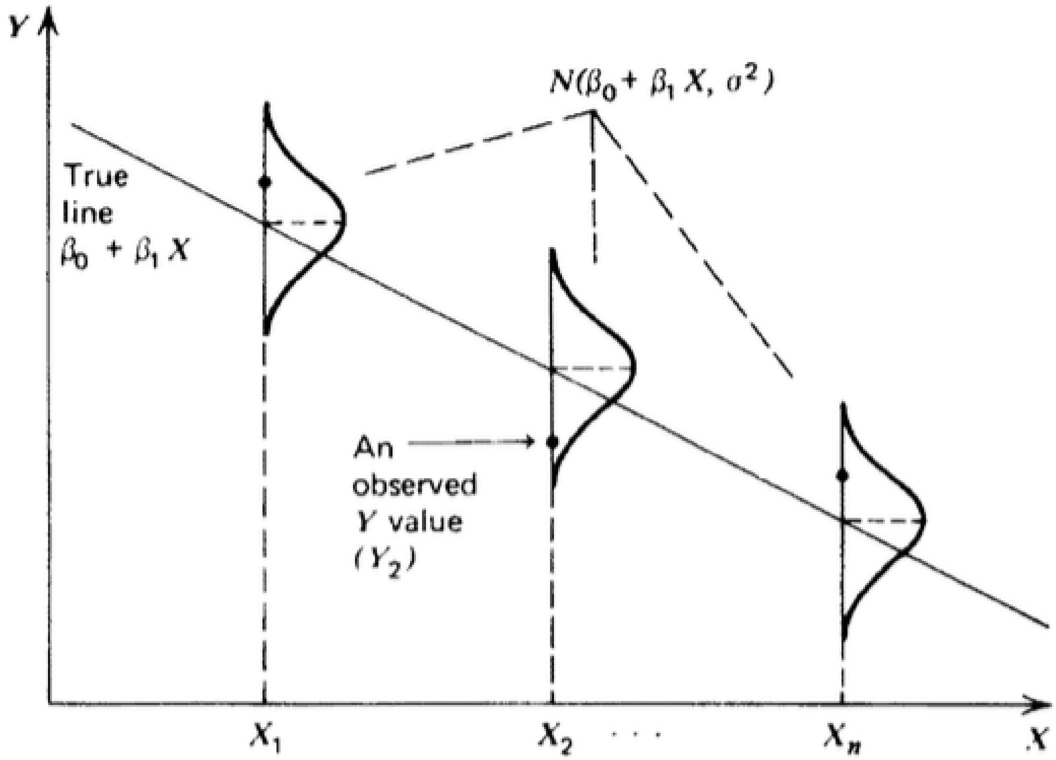
E a variância para a resposta Y_i é:

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

Utilizando a suposição 2, temos que a resposta Y_i vem de uma **distribuição Normal** com $E(Y_i) = \beta_0 + \beta_1 X_i$ (**função de regressão**) e $Var(Y_i) = \sigma^2$.

A resposta, Y_i está acima ou abaixo da função de regressão por um termo de erro ε_i que segue a **distribuição Normal**.

Propriedades de Y_i



Distribuição amostral de $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Suponha que tenhamos X_i's fixos, mas que observamos Y_i's várias vezes. A cada vez $\hat{\beta}_1$ será diferente.

$\hat{\beta}_1$ muda conforme mudamos nossa amostra. Desta forma, devemos estudar sua distribuição amostral.

A distribuição amostral de $\hat{\beta}_1$ dependerá das suposições que fizermos para o modelo de regressão.

Distribuição amostral de $\hat{\beta}_1$

Vimos que, sem suposição de distribuição de probabilidade, apenas com o momentos definidos, temos que:

$$E(\hat{\beta}_1) = \beta_1$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Se supomos que $\varepsilon_i \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, temos:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

Distribuição amostral de $\hat{\beta}_1$

Padronizando, temos que:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var(\hat{\beta}_1)}} \sim \mathcal{N}(0, 1)$$

Problema: $Var(\hat{\beta}_1)$ depende de σ^2, portanto é desconhecida.

Relembrando distribuição t-Student

Sejam $Z \sim \mathcal{N}(0,1)$ e $V \sim \chi^2_\nu$, **independentes**, então

$$T = \frac{Z}{\sqrt{V/\nu}} \sim t_\nu$$

Mostre[1] que

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var(\hat{\beta}_1)}}} \sim t_{n-2}$$

em que

$$\widehat{Var(\hat{\beta}_1)} = \frac{s^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad \text{e} \quad s^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$$

IMECC

Intervalos de confiança para os estimadores

Intervalo de confiança para β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})Y_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\widehat{Var(\hat{\beta}_1)} = \frac{s^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Um intervalo de $100(1-\alpha)\%$ de confiança para β_1 é dado por:

$$IC(\beta_1, 1-\alpha) = \left[\hat{\beta}_1 - t_{n-2,\alpha/2}\sqrt{\frac{s^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}};\right.$$

$$\left.\hat{\beta}_1 + t_{n-2,\alpha/2}\sqrt{\frac{s^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\right]$$

Intervalo de confiança para β_0

Especificamente, para o intercepto, temos que:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\widehat{Var(\hat{\beta}_0)} = s^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)$$

Intervalo de confiança para β_0

Um intervalo de $100(1-\alpha)\%$ de confiança para β_0 é dado por:

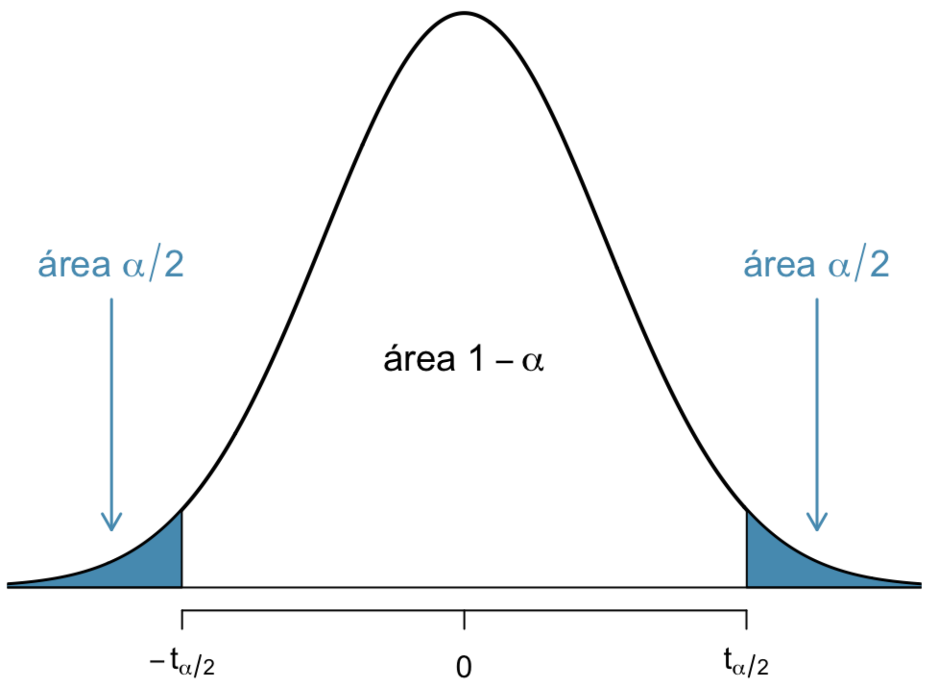
$$IC(\beta_0, 1-\alpha) = \left[\hat{\beta}_0 - t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)};\right.$$

$$\left.\hat{\beta}_0 + t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}\right]$$

Como encontrar $t_{n-2,\alpha/2}$

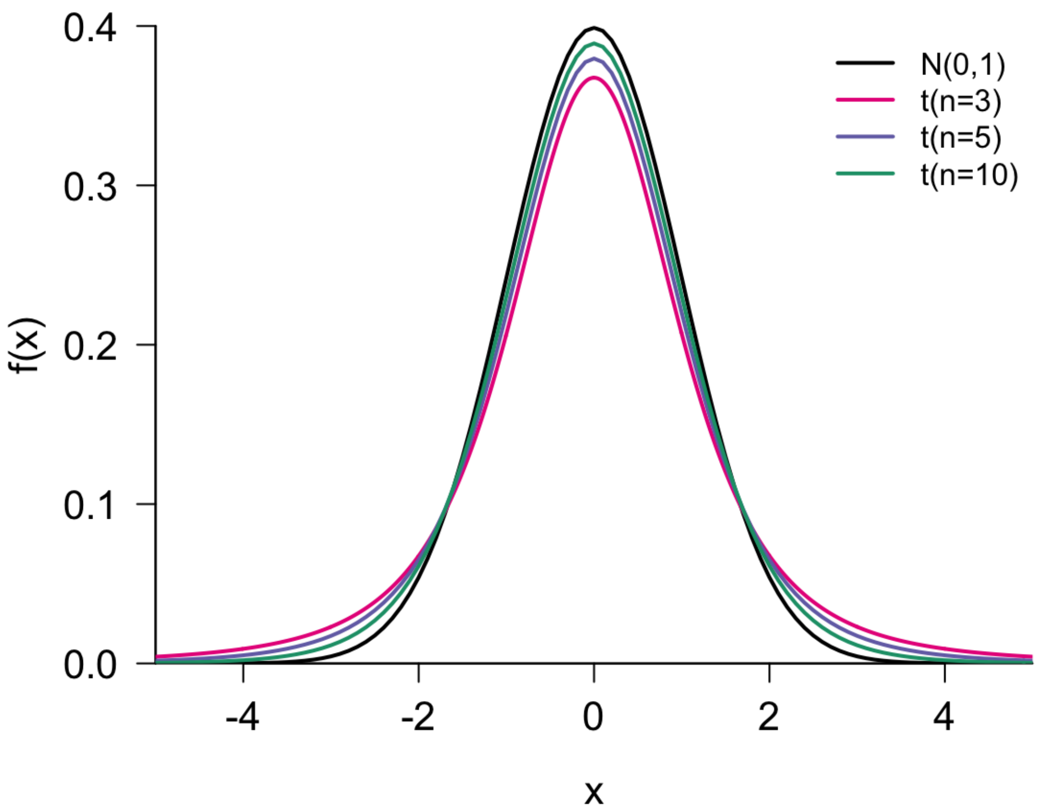
$$T \sim t_{n-2}$$

$$P(-t_{n-2,\alpha/2} < T < t_{n-2,\alpha/2}) = 1 - \alpha$$



Distribuição t-student e Normal Padrão

Para n grande a distribuição t-student se aproxima da normal padrão $N(0, 1)$.



Teste de Hipóteses

Revisão: Testes de Hipóteses

- Hipótese nula: H_0
- Hipótese alternativa: H_a
- Estatística do teste
- Nível de significância
- Região de rejeição

Revisão: Testes de Hipóteses

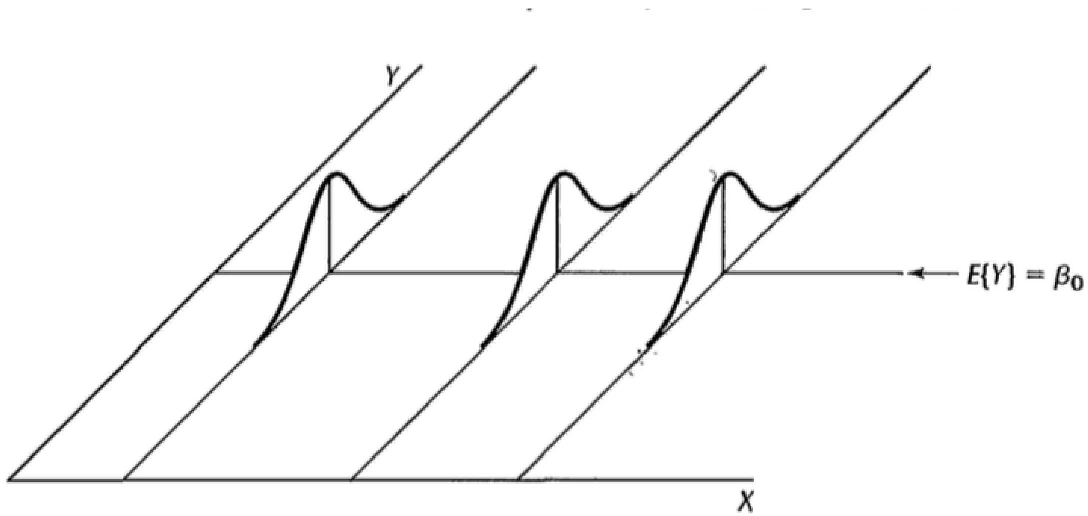
Erros:

- Erro do Tipo I: H_0 é rejeitada quando é verdadeira. A probabilidade do erro tipo I é α (nível de significância do teste).
- Erro do Tipo II: H_0 não é rejeitada quando H_a é verdadeira. A probabilidade do erro tipo II é β.

Decisão	H_0 é verdadeira	H_a é verdadeira
Aceita H_0	Correto	Erro Tipo II
Rejeita H_0	Erro Tipo I	Correto

Teste de Hipóteses para β_1

- $H_0\colon \beta_1 = 0$
- $H_a\colon \beta_1 \neq 0$



Teste de Hipóteses para β_1

Usaremos como estatística do teste nosso estimador: $\hat{\beta}_1$.

Se o valor observado, isto é, a estimativa $\hat{\beta}_1$ estiver longe de 0, temos evidências contra H_0.

Quão longe?

Temos que levar em conta a distribuição de probabilidade do estimador $\hat{\beta}_1$ quando H_0 é verdadeira:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var(\hat{\beta}_1)}}} \stackrel{H_0:\beta_1=0}{=} \frac{\hat{\beta}_1}{\sqrt{\widehat{Var(\hat{\beta}_1)}}} \stackrel{H_0:\beta_1=0}{\sim} t_{n-2}$$

Exemplo: Empresa Toluca

A empresa Toluca fabrica equipamentos de refrigeração e peças de reposição.

No passado, uma das peças de reposição era produzida periodicamente em lotes de tamanhos variável.

Para reduzir os custos, o diretor da empresa queria que se determinasse o tamanho ótimo do lote.

Para descobrir o tamanho ideal do lote é de extrema importância avaliar a relação entre tamanho do lote de total de horas trabalhadas na produção do mesmo.

Para tanto, avaliou-se o tamanho do lote e o número de horas para 25 lotes recentemente produzidos.

Exemplo: Empresa Toluca

$V1$: tamanho do lote

$V2$: horas trabalhadas para produzir o lote

10 primeiras observações do conjunto de dados:

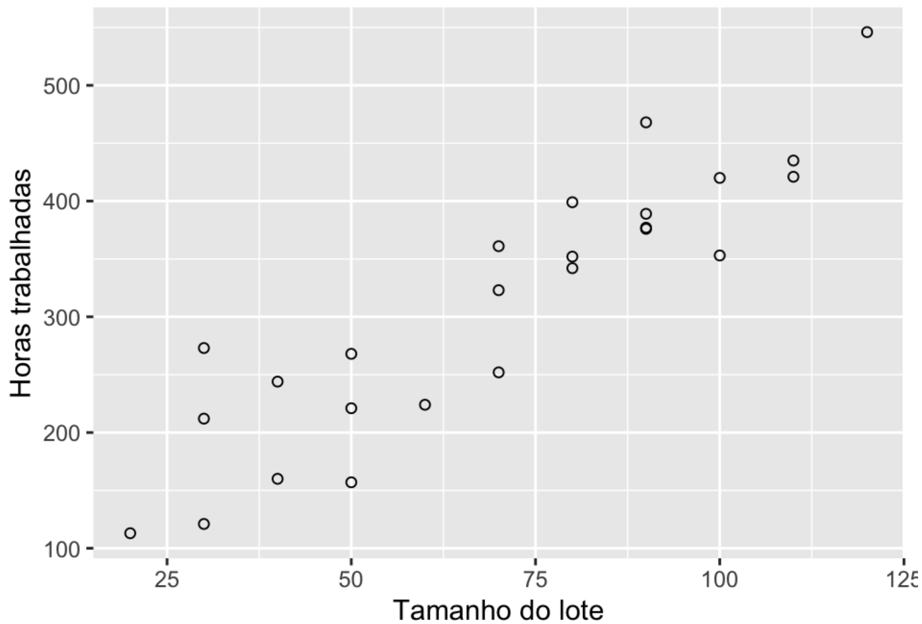
```
##      V1  V2
## 1    80 399
## 2    30 121
## 3    50 221
## 4    90 376
## 5    70 361
## 6    60 224
## 7   120 546
## 8    80 352
## 9   100 353
## 10   50 157
```

Exemplo: Empresa Toluca

Y_i: horas trabalhadas para produzir o lote i

X_i: tamanho do lote i

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Exemplo: Empresa Toluca

```
##
## Call:
## lm(formula = V2 ~ V1, data = dados)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.876 -34.088  -5.982  38.826 103.528
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.366     26.177   2.382   0.0259 *
## V1             3.570      0.347  10.290 4.45e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.82 on 23 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
## F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

Exemplo: Empresa Toluca

$\hat{\beta}_1 = 3.57$

$\hat{\beta}_0 = 62.37$

Estimamos que o número médio de horas trabalhadas aumenta em 3.57 para cada unidade adicional produzida no lote.

Esta estimativa se aplica a tamanhos de lote utilizados na análise: 20 a 120.

$\hat{E}(Y|X) = \hat{Y} = 62.37 + 3.57X.$

Desta forma, com a equação acima, podemos estimar o número esperado de horas trabalhadas para qualquer tamanho de lote.

Exemplo: Empresa Toluca

Por exemplo, segundo o modelo ajustado:

- para um lote de tamanho 30, temos que o número esperado de horas trabalhadas é 169.47.
- para um lote de tamanho 40, temos que o número esperado de horas trabalhadas é 205.17.

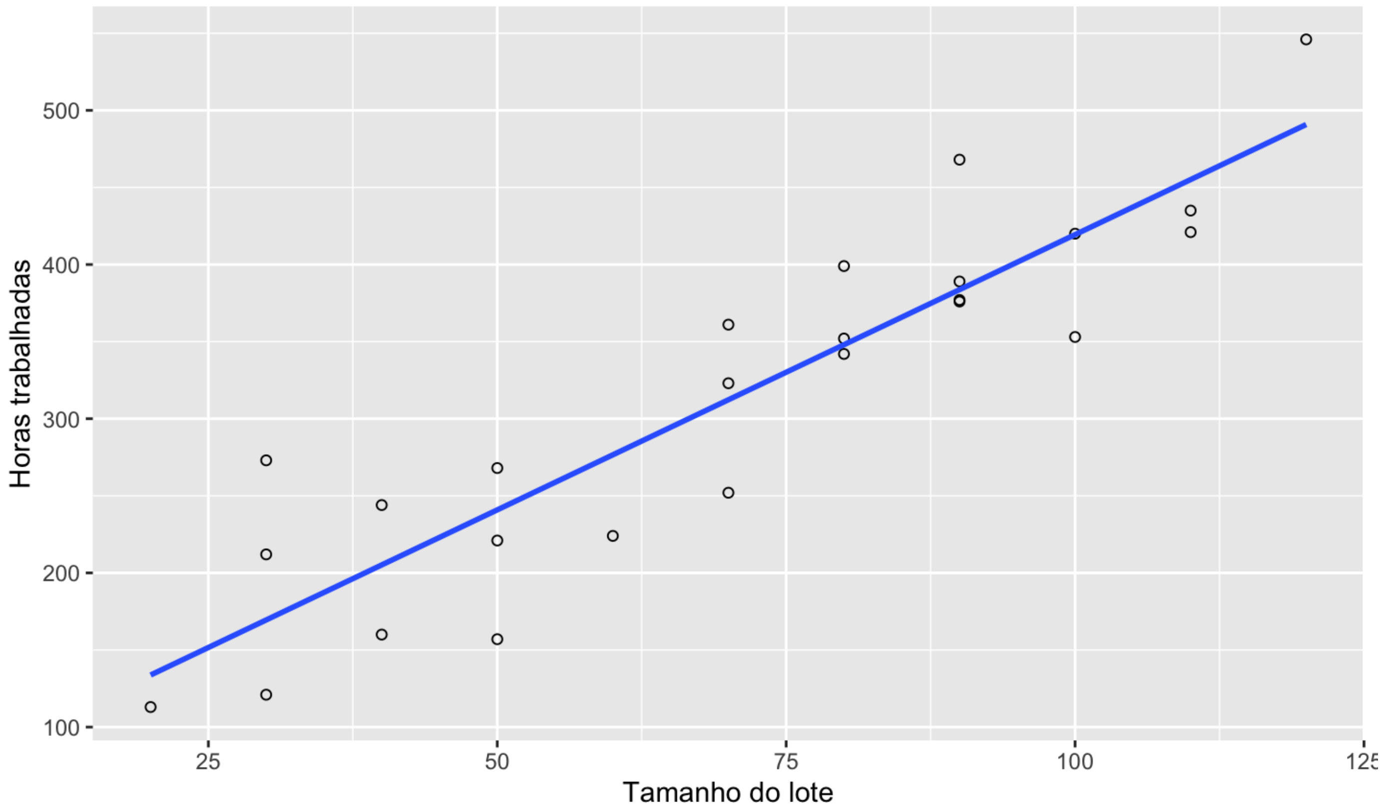
Observe que este valores são esperados e que há uma variabilidade ao redor desde valor.

Por exemplo: para lotes de tamanho 30, o valor esperado é 169.47, mas chegamos a observar 121, 212, 273 nos dados.

Para lotes de tamanho 40, o valor esperado é 205.17, mas chegamos a observar 160, 244 nos dados.

Exemplo: Empresa Toluca

```
## `geom_smooth()` using formula 'y ~ x'
```



Predição

Predição

A função ajustada, $\hat{Y}_i = \hat{E}(Y \mid X = X_i)$ pode ser usada para obter valores para a variável resposta para qualquer valor da variável preditora.

É importante distinguir dois problemas diferentes: predição e estimação de valores ajustados.

Em predição, temos uma nova observação cujo valor da variável preditora é X^*, possivelmente uma observação futura, não utilizada nas estimativas dos parâmetros.

Queremos saber o valor de Y^* correspondente, mas que ainda não foi observado.

Predição

Podemos usar a reta estimada de maneira que uma predição pontual de Y^*, indicada por $\tilde{Y}^*$ é dada por:

$$\tilde{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X_i^*$$

Quão precisa é esta estimativa?

$$Var(\tilde{Y}^* \mid X = X_i^*) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(X_i^* - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)$$

(ver seção 2.6.3)

Exemplo: precificação de diamantes

```
library(UsingR); data(diamond)
y <- diamond$price; x <- diamond$carat; n <- length(y)
beta1 <- cor(y, x) * sd(y) / sd(x)
beta0 <- mean(y) - beta1 * mean(x)
e <- y - beta0 - beta1 * x
sigma <- sqrt(sum(e^2) / (n-2))
ssx <- sum((x - mean(x))^2)
seBeta0 <- (1 / n + mean(x) ^ 2 / ssx) ^ .5 * sigma
seBeta1 <- sigma / sqrt(ssx)
tBeta0 <- beta0 / seBeta0; tBeta1 <- beta1 / seBeta1
pBeta0 <- 2 * pt(abs(tBeta0), df = n - 2, lower.tail = FALSE)
pBeta1 <- 2 * pt(abs(tBeta1), df = n - 2, lower.tail = FALSE)
coefTable <- rbind(c(beta0, seBeta0, tBeta0, pBeta0), c(beta1, seBeta1, tBeta1, pBeta1))
colnames(coefTable) <- c("Estimate", "Std. Error", "t value", "P(>|t|)")
rownames(coefTable) <- c("(Intercept)", "x")
```

Exemplo: precificação de diamantes

```
coefTable
```

```
##                Estimate Std. Error   t value      P(>|t|)
## (Intercept) -259.6259   17.31886 -14.99094 2.523271e-19
## x           3721.0249   81.78588  45.49715 6.751260e-40
```

```
fit <- lm(y ~ x);
summary(fit)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -259.6259   17.31886 -14.99094 2.523271e-19
## x           3721.0249   81.78588  45.49715 6.751260e-40
```

Exemplo: precificação de diamantes

IC para o intercepto:

```
sumCoef <- summary(fit)$coefficients
sumCoef[1,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef[1, 2]
```

```
## [1] -294.4870 -224.7649
```

Equivalente:

```
confint(fit)[1,]
```

```
##      2.5 %    97.5 %
## -294.4870 -224.7649
```

Exemplo: precificação de diamantes

IC para o aumento de preço para 0.1 carat de aumento no tamanho do diamante:

```
(sumCoef[2,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef[2, 2]) / 10
```

```
## [1] 355.6398 388.5651
```

Equivalente:

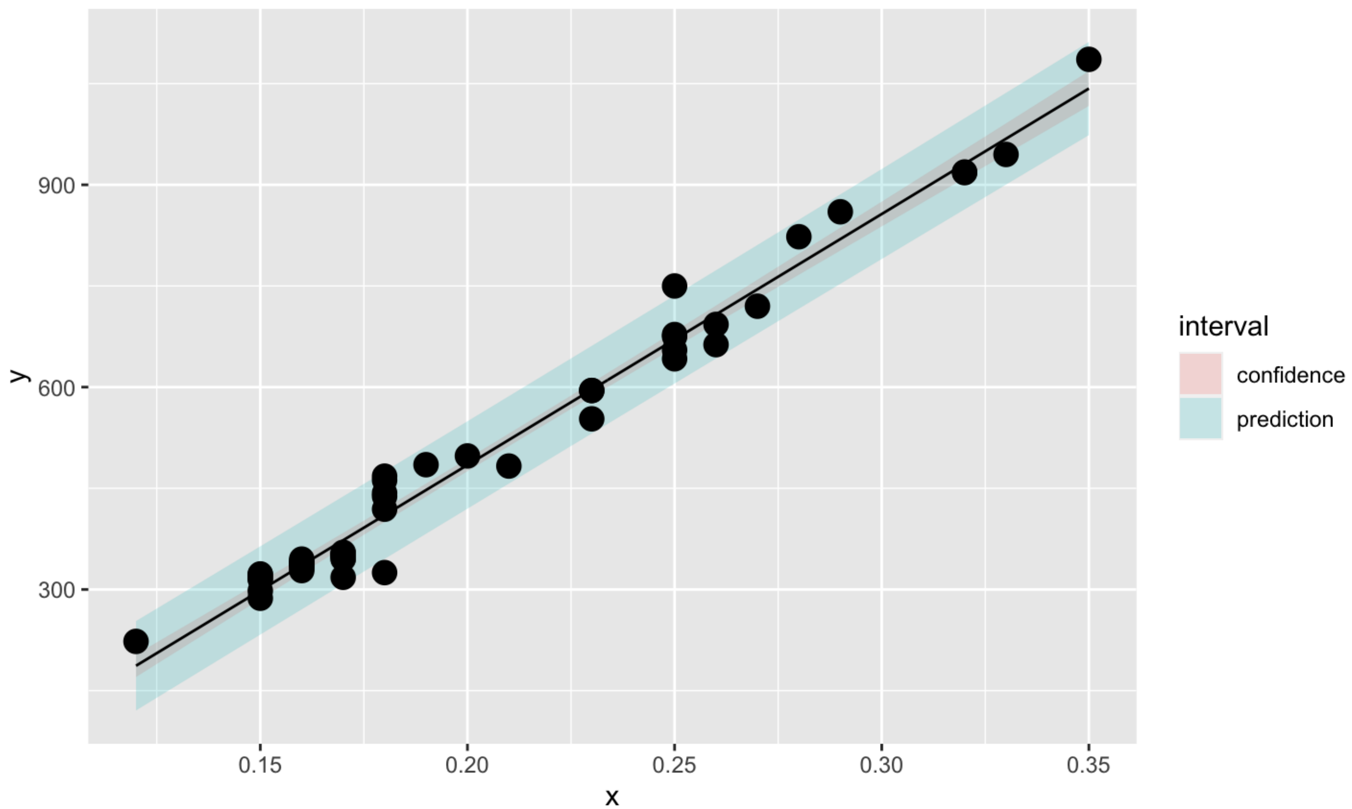
```
confint(fit)[2,]/10
```

```
##    2.5 %   97.5 %
## 355.6398 388.5651
```

Com 95% de confiança, estimamos que um aumento de 0.1 carat no tamanho do diamante resulta em um aumento médio entre 355.6 e 388.6 dólares.

Exemplo: precificação de diamantes

Agora, vamos verificar a precisão quando fazemos predições.



Discussão

Lembrem que tínhamos dois problemas diferentes: predição e estimação de valores ajustados.

- Em rosa temos o intervalo para a reta estimada (valor esperado).
- Em azul temos o intervalo para predição de valores pontuais.
- Ambos os intervalos têm comprimento variável para diferentes valores de X.
- Ambos os intervalos apresentam menor comprimento perto de $\bar{X}$.
- A confiança na reta ajustada é maior, portanto seu IC tem menor comprimento
- O intervalo para a predição incorpora a variabilidade dos dados ao redor da reta.

Exemplo: usando a função `predict` do R

```
x <- rnorm(15)
y <- x + rnorm(15)
predict(lm(y ~ x))
```

```
##           1           2           3           4           5           6
## -0.86832583  0.18908891  0.19365522  1.48712024 -0.33863165  0.03040501
##           7           8           9          10          11          12
##  0.95527061  0.31123295 -0.18949875  1.23741838  0.93969768  0.37310015
##          13          14          15
## -0.04617044 -0.17123272 -0.36746172
```

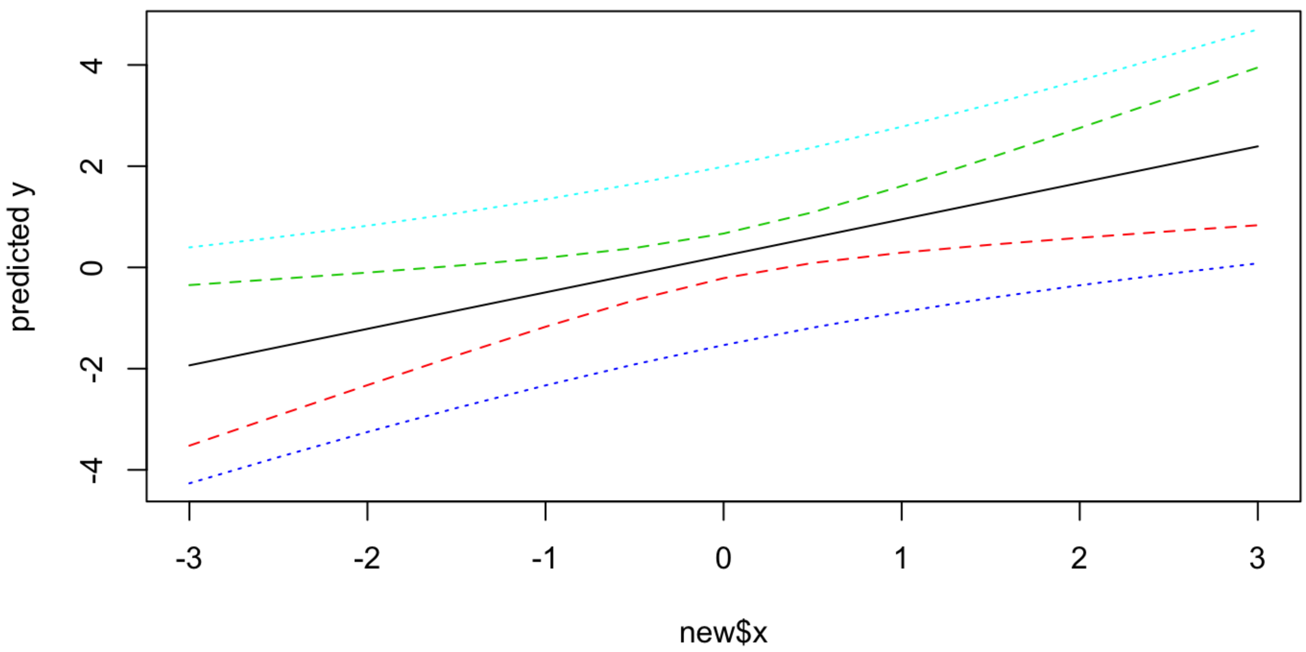
Exemplo: usando a função `predict` do R

```r
new <- data.frame(x = seq(-3, 3, 0.5))
predict(lm(y ~ x), new, se.fit = TRUE)
```

```
## $fit
##          1          2          3          4          5          6          7
## -1.9352156 -1.5746903 -1.2141650 -0.8536397 -0.4931144 -0.1325891  0.2279362
##          8          9         10         11         12         13
##  0.5884615  0.9489868  1.3095121  1.6700374  2.0305626  2.3910879
##
## $se.fit
##         1         2         3         4         5         6         7         8
## 0.7342326 0.6232424 0.5146552 0.4103831 0.3147442 0.2383672 0.2041549 0.2316137
##         9        10        11        12        13
## 0.3044930 0.3986130 0.5021694 0.6103779 0.7211475
##
## $df
## [1] 13
##
## $residual.scale
## [1] 0.7902479
```

Exemplo: usando a função `predict` do R

```
pred.w.plim <- predict(lm(y ~ x), new, interval = "prediction")
pred.w.clim <- predict(lm(y ~ x), new, interval = "confidence")
matplot(new$x, cbind(pred.w.clim, pred.w.plim[,-1]),
        lty = c(1,2,2,3,3), type = "l", ylab = "predicted y")
```



Leituras sobre suposição de normalidade

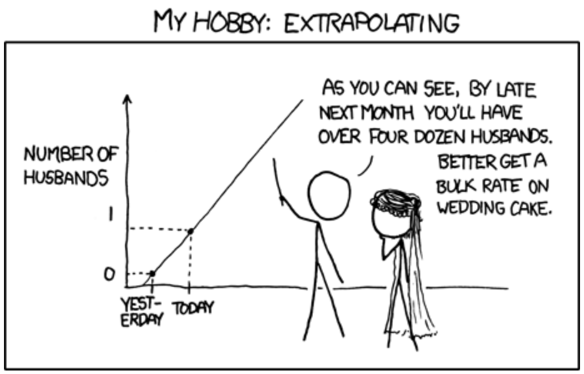
- The Importance of the Normality Assumption in Large Public Health Data Sets
- The Assumption(s) of Normality
- Understanding Regression Assumptions

Agradecimento

- Slides criados por Samara F Kiihl / IMECC / UNICAMP

Leitura

- Applied Linear Statistical Models: 1.8, 2.1-2.9.
- Caffo - Regression Models for Data Science in R: Regression Inference.
- Draper & Smith - Applied Regression Analysis: 1.4, 1.5, 3.1.
- Weisberg - Applied Linear Regression: Capítulo 2.



xkcd