Métodos Quantitativos

Aula 09. Regressão linear, parte 1

Pedro H. G. Ferreira de Souza

pedro.ferreira@ipea.gov.br

Mestrado Profissional em Políticas Públicas e Desenvolvimento

Instituto de Pesquisa Econômica Aplicada (Ipea)

21 nov. 2022

O que vimos até aqui

Aulas 1 e 2

- Metodologia de pesquisa e causalidade

Aulas 3 a 5

- Introdução à manipulação de dados no R; estatísticas descritivas uni- e bivariadas

Aula 6

- Amostragem, variáveis aleatórias e distribuições amostrais

Aulas 7 e 8

- Intervalos de confiança
- Testes de hipóteses
- Comparações entre médias

Perguntas

Vimos gráficos, médias condicionais, correlações...mas ainda temos perguntas importantes que não respondemos:

1. Qual a mudança em Y se X varia?
2. Qual o efeito de X em Y se controlarmos (ou seja, tirarmos o efeito) de outras variáveis?
3. Como prever valores de Y a partir de X, Z e outras?

Para responder, precisamos aprender **análise de regressão**.

- Vamos pressupor forma funcional linear
- Serve para objetivos descritivos, causais ou preditivos

Distribuições e médias condicionais

Distribuições marginais

- São as distribuições das variáveis tomadas individualmente, com suas médias, desvios padrão etc; é o que obtemos com tabelas de frequência, histogramas etc.
- E.g: a probabilidade marginal de uma pessoa ser do sexo masculino é de 50% $\rightarrow Pr(Masculino) = 50\%$.

Distribuições condicionais

- São distribuições de uma variável condicionadas a um valor fixo de outra variável; é o que obtemos com tabelas cruzadas, médias condicionais etc.
- E.g.: a probabilidade de uma pessoa *chamada Alcione* ser do sexo masculino é de 91% $\rightarrow Pr(Masculino \mid Nome = Alcione) = 23\%$.

Esperanças condicionais

Exemplo #1. Suponha que jogamos dois dados, x e z. Sabemos que $x = 4$, mas não sabemos z ainda. Qual o valor esperado para a soma dos dois?

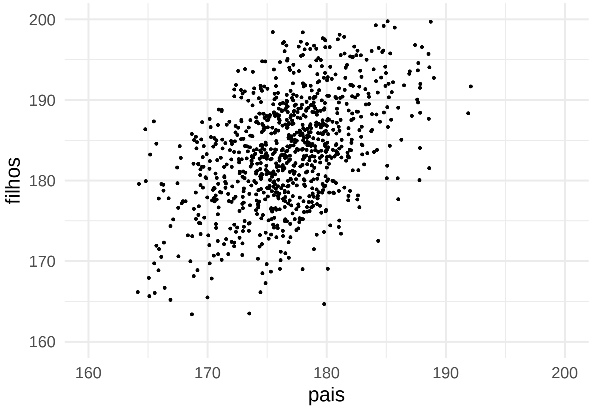
$$E(x + z \mid x = 4) = 7.5$$

Exemplo #2. Qual o valor esperado para o número de sobrinhos para pessoas com 0, 1, 2... irmãos?

$$E(S \mid I = 0) = 0, \qquad E(S \mid I = 1) = y_1, \qquad E(S \mid I = 2) = y_2, \qquad \ldots$$

Exemplo #3. Qual o valor esperado para a altura de crianças condicional à altura dos pais?

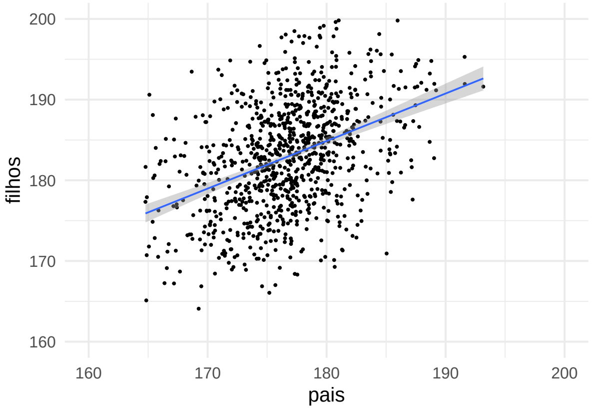
Exemplo simulado para altura (i)




```
cor(df$pais, df$filhos)
```

```
## [1] 0.4773441
```

Exemplo simulado para altura (ii)




```
lm(df$filhos ~ df$pais)$coefficients
```

```
## (Intercept)     df$pais
##  74.2879544   0.6154085
```

Pressupostos

Vamos pressupor **linearidade** na relação entre X e Y porque é bem mais simples & bastante flexível

- Se o modelo populacional não for linear, o R vai rodar a regressão, mas os resultados serão inúteis

A linearidade é só um dos pressupostos que veremos para que o modelo funcione bem.

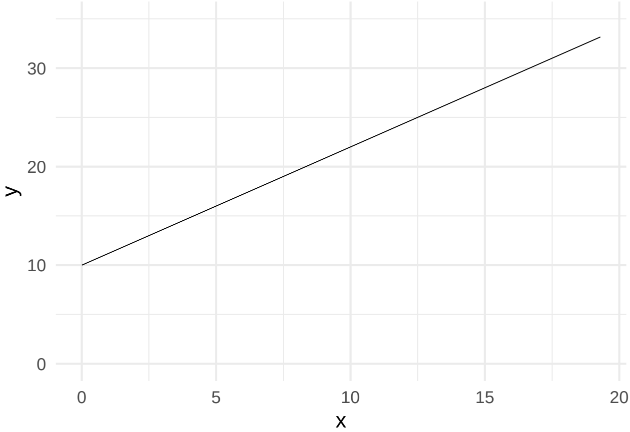
A interpretação causal dos coeficientes também depende dos pressupostos

- Causalidade: pressupostos + modelo + dados.

Pacotes e dados

```
# Pacotes
library(tidyverse)
library(HistData)
# Dados
galton.df <- GaltonFamilies %>%
              filter(!is.na(childHeight) &
                       !is.na(midparentHeight)) %>%
              mutate(filhos = 2.54 * childHeight,
                     pais = 2.54 * midparentHeight,
                     mae = 2.54 * mother,
                     pai = 2.54 * father)
```

Função linear determinística, $y = \alpha + \beta x$




```
## [1] Y = 10 + 1.2*x
```

Modelos e realidade

Modelos são sempre **simplificações úteis** da realidade que usamos para descrever e explicar padrões.

O mundo social não é determinístico, mas **probabilístico**. Há sempre incerteza devido à aleatorização, variáveis omitidas, choques aleatórios etc.

- Nenhum *scatter plot* de dados reais jamais vai ser como o gráfico do slide anterior…

Ao introduzir *outras causas* + *erros aleatórios* no modelo, estimamos:

$$y = \alpha + \beta x + \epsilon \qquad \text{ou seja,} \qquad E(y \mid x) = \alpha + \beta x$$

O que significa $y = \alpha + \beta x + \epsilon$?

- Estamos pressupondo que no **universo** dois fenômenos de interesse têm relação linear não determinística.
- Em geral, vamos usar uma **amostra** para estimar os parâmetros de interesse, $\hat{\alpha}$ e $\hat{\beta}$ (ou a e b)
- A linearidade no universo é um **pressuposto**
  - Inclinação não muda conforme o valor de x, isto é, $\frac{\delta y}{\delta x} = c$
  - Se nossa amostra for representativa, deve refletir a linearidade (ou a falta dela) na população. Por isso, a **inspeção visual** via *scatter plots* é uma etapa inicial **indispensável**

O que significa $y = \alpha + \beta x + \epsilon$?

Esperança condicional

$$E(y \mid x) = \alpha + \beta x$$

$$y_i = E(y \mid x = x_i) + \epsilon_i = \alpha + \beta x_i + \epsilon_i$$

O estimador mais simples nesse caso é o de **mínimos quadrados ordinários** (MQO ou OLS, em inglês).

Pressupostos do modelo clássico:

1. Linearidade nos coeficientes
2. Exogeneidade estrita: $E(\epsilon \mid x) = 0$
3. Variância esférica dos erros: $var(\epsilon \mid x) = \sigma_e^2$
4. Erros não correlacionados
5. Independência linear entre os x

Recapitulação

Introdução

Relações lineares

Estimativas de ponto por MQO

Ajuste do modelo

Inferência e testes de hipóteses

Resumo da aula

Próxima aula

O estimador MQO

MQO estima parâmetros populacionais minimzando a **soma dos quadrados dos erros**.

- Dados alguns pressupostos, ele é ótimo: pelo teorema de Gauss-Markov, é **BLUE** (estimador linear não enviesado mais eficiente)
- Que erros são esses? Lembre-se que estamos simplificando o mundo em uma relação linear, "passando uma reta". Logo, não prevemos com exatidão cada observação.

Exemplo bobo

Suponha que temos uma variável contínua X. Quero resumir essa variável em único número. Qual seria? Pelo critério de MQO, o melhor palpite é a **média**.

Estimação: $y_i = \alpha + \beta x_i + \epsilon_i$

Cada y é a soma de um componente sistemático $\hat{y} = \alpha + \beta x_i$ e de um resíduo aleatório ϵ_i.

MQO estima os valores que minimizam a **soma dos quadrados dos erros**. Como $\epsilon_i = y_i - (\alpha + \beta x_i)$, queremos minimizar:

$$SQ(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Os valores estimados a e b são obtidos por:

$$a = \bar{y} - b\bar{x} \qquad b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{cov(x, y)}{var(x)} = corr(x, y)\frac{sd(y)}{sd(x)}$$

Por que funciona?

Como sabemos que o estimador é **não enviesado**, isto é, $E(b) = \beta$, se não observamos ϵ_i? Precisamos de dois pressupostos fundamentais.

Pressupostos para ausência de viés

1. Linearidade
   - Forma funcional linear na população
2. Exogeneidade estrita
   - Erro ortogonal a x: $E(\epsilon) = E(\epsilon \mid x) = 0 \rightarrow cov(x, \epsilon) = 0$
   - Pressuposto essencial para afirmações causais, garantido somente com aleatorização
   - Violações comuns: simultaneidade, variáveis omitidas, erros de medida etc $\rightarrow b = \beta + \frac{cov(x,\epsilon)}{var(x)}$

Estimação de ponto na prática

```
# MQO manual
galton.df %>% summarise(b = cov(filhos, pais) / var(pais),
                        a = mean(filhos) - b*mean(pais))
# MQO automatico
lm(filhos ~ pais, data = galton.df)
```

```
##           b        a
## 1 0.6373609 57.49605
##
## Call:
## lm(formula = filhos ~ pais, data = galton.df)
##
## Coefficients:
## (Intercept)         pais
##     57.4961       0.6374
```

Os coeficientes

O coeficiente b

Para variáveis em nível, b reflete a variação em y associada a uma mudança em uma unidade em x.

O intercepto a

Se o modelo inclui o intercepto, a regressão explica variações em torno das médias. Como $a = \bar{y} - b\bar{x}$:

$$y_i = (\bar{y} - b\bar{x}) + bx_i + e_i \qquad \rightarrow \qquad y_i - \bar{y} = b(x_i - \bar{x}) + e_i$$

O intercepto é o valor predito quando $x = 0$. Em geral, não é muito importante: seu papel é servir como "coletor de lixo" de vieses não incorporados no modelo, garantindo que $\sum_{i=1}^{n} e_i = 0$.

Observações

O ponto $(\bar{x}, \bar{y})$

Por definição, a inclusão do intercepto faz com que $\bar{y} = a + b\bar{x}$, ou seja, o resíduo é igual a zero quando y e x têm valor médio.

Efeito de transformações lineares

Se transformarmos y em $zy + c$...

- ...os coeficientes vão se alterar para $(az + c)$ e bz.

Se transformamos x em $zx + c$...

- ...os coeficientes vão se alterar para $(a - bc)$ e b/z.

Exercício

O que acontece se normalizarmos y e x, isto é, se estimarmos com $(y - \bar{y})/sd(y)$ e $(x - \bar{x})/sd(x)$?

Exercício

O que acontece se normalizarmos y e x, isto é, se estimarmos com $(y - \bar{y})/sd(y)$ e $(x - \bar{x})/sd(x)$?

```
galton.df <-  galton.df %>%
    mutate(filhos_zsc = (filhos - mean(filhos)) / sd(filhos),
           pais_zsc = (pais - mean(pais)) / sd(pais))
```

Exercício

O que acontece se normalizarmos y e x, isto é, se estimarmos com $(y - \bar{y})/sd(y)$ e $(x - \bar{x})/sd(x)$?

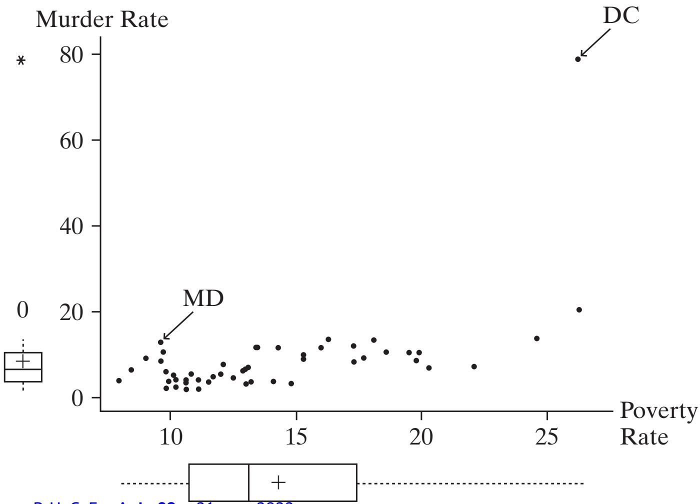
```
galton.df <-  galton.df %>%
    mutate(filhos_zsc = (filhos - mean(filhos)) / sd(filhos),
          pais_zsc = (pais - mean(pais)) / sd(pais))

mqo_zsc <- lm(filhos_zsc ~ pais_zsc, data = galton.df)
mqo_zsc$coefficients %>% round(digits = 7)
cor(galton.df$filhos_zsc, galton.df$pais_zsc)
```

```
## (Intercept)    pais_zsc
##   0.0000000   0.3209499
## [1] 0.3209499
```

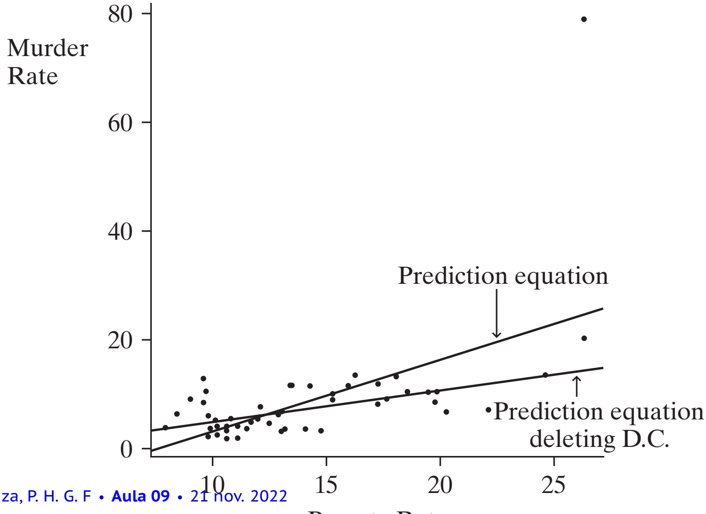
Outliers (i)

Agresti, 2018, Figura 9.3



Outliers (ii)

Agresti, 2018, Figura 9.5



Outliers (iii)

O que fazer?

- Inspeção visual uni- e bivariada: há *outliers*?
  - Erro de medida ou problema real?
- Realizar testes estatísticos para *leverage* e *influence*
  - DFBETA, DFFITS, D de Cook etc
- Na prática, problema tende a ser muito maior em amostras pequenas & quando não há limites "naturais" para y ou x
  - No exemplo de Galton, é impossível termos alguém com 10cm ou 10m de altura...
  - ... mas podemos sortear por acaso um bilionário ou um *influencer* etc

Exercício

O data frame `Cholera`, do pacote `HistData` contém dados sobre a epidemia de cólera nos distritos de Londres em 1849.

- `cholera_drate` é a taxa de mortes por 10,000 habitantes
- `elevation` é a elevação acima do nível do Rio Tâmisa em pés

Quais os coeficientes da regressão de `cholera_drate` sobre `elevation`?

Exercício

O data frame `Cholera`, do pacote `HistData` contém dados sobre a epidemia de cólera nos distritos de Londres em 1849.

- `cholera_drate` é a taxa de mortes por 10,000 habitantes
- `elevation` é a elevação acima do nível do Rio Tâmisa em pés

Quais os coeficientes da regressão de `cholera_drate` sobre `elevation`?

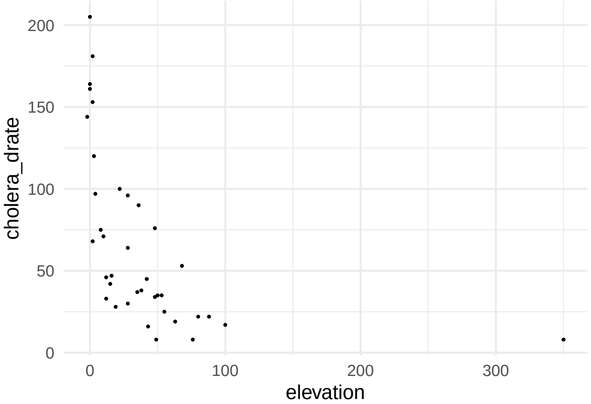
```
lm(cholera_drate ~ elevation, data = Cholera)
```

```
##
## Call:
## lm(formula = cholera_drate ~ elevation, data = Cholera)
##
## Coefficients:
## (Intercept)    elevation
##     83.8117      -0.4388
```

Exercício

Reparem na presença de grandes *outliers*:

```
Cholera %>% ggplot(aes(x = elevation, y = cholera_drate)) +
    geom_point() + theme_minimal(base_size = 24)
```



Exercício

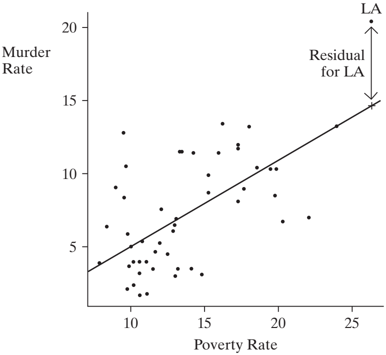
Reparem na presença de grandes *outliers*:

```
colera <- Cholera %>% filter(elevation < 300)
lm(cholera_drate ~ elevation, data = colera)
```

```
##
## Call:
## lm(formula = cholera_drate ~ elevation, data = colera)
##
## Coefficients:
## (Intercept)    elevation
##     110.004       -1.325
```

Valores preditos e resíduos

- O **valor observado** é y_i
- O **valor predito** é $\hat{y}_i = a + bx_i$
- O **resíduo** é $e_i = y_i - \hat{y}_i$
  - No modelo com intercepto, por definição, $\bar{e}_i = 0$ e $\bar{y} = \bar{\hat{y}}$



Decomposição da soma dos quadrados

Por definição, MQO minimiza a **soma dos quadrados dos resíduos**, isto é, $SSE = \sum(y_i - \hat{y}_i)^2$.

Observe que $SST = SQR + SSE$, com:

- SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$ *(soma total dos quadrados)*
- SSE = $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ *(soma dos quadrados dos erros)*
- SSR = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ *(soma dos quadrados do modelo)*

Afinal, temos:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) = e_i + (\hat{y}_i - \bar{y})$$

Estatísticas de ajuste

r^2 ou coeficiente de determinação

É a proporção da variância de y "explicada" pelo modelo:

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Em regressões univariadas com intercepto, o r^2 é o quadrado do coeficiente de correlação de Pearson.

$$r_{xy} = \frac{cov(x, y)}{sd(x)sd(y)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Sobre o r^2

- O r^2 não depende de unidade de mensuração e, como a correlação, mede apenas a força da associação linear.
  - Como $-1 \leq r_{xy} \leq 1$, o r^2 também fica entre 0 e 1 (mas quando não incluímos o intercepto, o r^2 pode assumir valores negativos)
  - Para que $r^2 = 1$, é preciso que $SSE = 0$, ou seja, que todos os pontos caiam exatamente na linha da regressão
- O r^2 só é comparável entre modelos quando eles são para a mesma variável dependente.
- O r^2 ser baixo não é necessariamente um problema.
  - Nosso objetivo quase nunca é explicar a maior parte da variância de y, mas sim avaliar o efeito de x.
  - O r^2 não diz nada sobre a qualidade das estimativas populacionais.

r^2 no R

```
modelo <- lm(cholera_drate ~ elevation, data = Cholera)
resumo <- summary(modelo)
print(resumo$r.squared)
```

```
## [1] 0.232628
```

```
cor(Cholera$cholera_drate, Cholera$elevation)^2
```

```
## [1] 0.232628
```

```
var(modelo$fitted.values) / var(Cholera$cholera_drate)
```

```
## [1] 0.232628
```

Erro padrão da regressão

O **erro padrão da regressão** (σ_ϵ) quantifica (na escala da variável y) o tamanho médio dos erros. Um estimador não viesado é:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-p}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}} = \sqrt{\frac{SSE}{n-p}}$$

Em que:

- n é o tamanho da amostra
- p é o número de parâmetros (no caso, α e β)

Observação: estimador não viesado somente com o pressuposto de ausência de autocorrelação entre os erros.

Erro padrão da regressão no R

```
modelo <- lm(filhos ~ pais, data = galton.df)
resumo <- summary(modelo)

# Manualmente
sqrt(sum(resumo$residuals^2) / (nrow(galton.df) - 2))

## [1] 8.614952

# Automaticamente
print(resumo$sigma)

## [1] 8.614952
```

Resumo para estimativas de ponto

Ausência de viés

Em uma amostra aleatória, o modelo $y_i = E(y \mid x_i) + \epsilon_i = \alpha + \beta x_i + \epsilon_i$ deve satisfazer:

1. Linearidade
2. Exogeneidade estrita

$$E(a) = \alpha \qquad\qquad E(b) = \beta$$

Observe que em uma amostra aleatória $cov(\epsilon_i, \epsilon_j) = 0$ para $i \neq j$, isto é, para duas observações distintas, os erros não são correlacionados.

Variâncias dos estimadores por MQO

Pressuposto adicional

- Homoscedasticidade, isto é, variância constante: $var(\epsilon \mid x) = \sigma^2_\epsilon$

Com isso, temos $var(y_i \mid x_i) = \sigma^2_\epsilon$, de modo que:

$$var(a) = \frac{\sigma^2_\epsilon}{n} \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2_\epsilon}{n} \left(\frac{\overline{x_i^2}}{var(x)} \right) \qquad var(b) = \frac{\sigma^2_\epsilon}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2_\epsilon}{n} \frac{1}{var(x)}$$

Para σ^2_ϵ, usamos o estimador s^2_e:

$$s^2_e = \frac{\sum_{i=1}^{n} e_i^2}{n - p} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - p} = \frac{SSE}{n - p}$$

Ausência de viés + eficiência

Pressupostos

1. Linearidade nos parâmetros $\rightarrow y_i = \alpha + \beta x_i + \epsilon_i$
2. Exogeneidade estrita, isto é, média condicional do erro é $E(\epsilon \mid x) = 0$
3. Sem colinearidade perfeita $\rightarrow x$ não é uma constante
4. Amostragem aleatória $\rightarrow cov(\epsilon_i, \epsilon_j) = 0$
5. Homoscedasticidade $\rightarrow var(\epsilon \mid x) = \sigma_\epsilon^2$

Teorema de Gauss-Markov

Dado 1-5, os estimadores de MQO são os **melhores estimadores lineares não viesados (BLUE)**.

- "melhor" $\rightarrow$ mais eficientes == menor variância entre estimadores não viesados
- "estimador" $\rightarrow$ regra aplicada a uma amostra para estimar parâmetros populacionais
- "linear" $\rightarrow$ o estimador é uma função linear dos dados
- "não viesado" $\rightarrow E(b) = \beta$

Distribuição amostral dos estimadores

Para podermos construir ICs e fazermos testes de hipóteses, só falta sabermos a **distribuição amostral** dos estimadores.

- Em **amostras grandes**, não precisamos de nenhum pressuposto adicional (consistência + normalidade assintótica)
- Em **amostras pequenas** precisamos do pressuposto de que o erro populacional tem distribuição $\epsilon \sim N(0, \sigma_\epsilon^2)$

$$y_i \mid x_i \sim N(\alpha + \beta x_i, \sigma_\epsilon^2)$$

$$a \sim N\left(\alpha, \frac{s_e^2}{n}\left(1 + \frac{\bar{x}^2}{var(x)}\right)\right) \qquad b \sim N\left(\beta, \frac{s_e^2}{n}\frac{1}{var(x)}\right)$$

ICs e testes de hipótese

Tudo isso significa que podemos construir ICs e fazer testes para os coeficientes de forma parecida com o que já vimos:

Erro padrão de b

$$se_b = \frac{s_e}{\sqrt{n} sd(x)}, \qquad \text{com} \qquad s_e = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Intervalo de confiança para b

$$b \pm t_{n-2} se_b$$

Testes de hipóteses para b

$$\frac{b - \beta_{H0}}{} \sim t_{n-2}$$

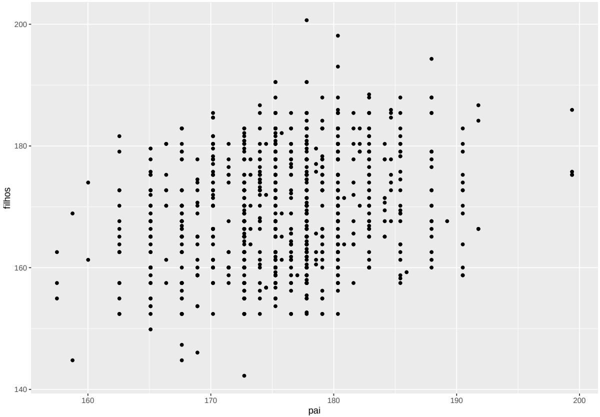
Exemplo/exercício

Em `galton.df`, faça o *scatter plot* da altura dos filhos (eixo vertical) sobre a altura do pai (eixo horizontal). A relação parece linear?

Exemplo/exercício

Em `galton.df`, faça o *scatter plot* da altura dos filhos (eixo vertical) sobre a altura do pai (eixo horizontal). A relação parece linear?

```
qplot(x = pai, y = filhos, data = galton.df, geom = 'point')
```



Aplicação prática

Exemplo/exercício (i)

Em `galton.df`, estime regressão da altura dos filhos sobre a altura somente do pai. Interprete os coeficientes.

Exemplo/exercício (i)

Em `galton.df`, estime regressão da altura dos filhos sobre a altura somente do pai. Interprete os coeficientes.

```
mod_pai <- lm(filhos ~ pai, data = galton.df)
print(mod_pai)
```

```
##
## Call:
## lm(formula = filhos ~ pai, data = galton.df)
##
## Coefficients:
## (Intercept)          pai
##    101.9538       0.3845
```

Exemplo/exercício (ii)

Qual o r^2 e qual o erro padrão da regressão? Interprete.

Exemplo/exercício (ii)

Qual o r^2 e qual o erro padrão da regressão? Interprete.

```
resumo_pai <- summary(mod_pai)
print(resumo_pai$r.squared)
```

```
## [1] 0.0707765
```

```
print(resumo_pai$sigma)
```

```
## [1] 8.76837
```

Exemplo/exercício (iii)

Qual o erro padrão de b? Inteprete.

Exemplo/exercício (iii)

Qual o erro padrão de *b*? Inteprete.

```
# Manualmente
s_e <- sqrt(sum(mod_pai$residuals^2) / (mod_pai$df.residual))
n <- mod_pai$rank + mod_pai$df.residual
sd_x <- sd(mod_pai$model$pai)
se_b <- s_e / (sqrt(n-1) * sd_x) # observem o n-1
print(se_b)
```

```
## [1] 0.04563621
```

```
# Automatico
resumo_pai$coefficients[1:2,1:2]
```

```
##                Estimate Std. Error
## (Intercept) 101.953809 8.02618067
## pai           0.384505 0.04563621
```

Exemplo/exercício (iv)

Qual o IC a 95% de b? Inteprete.

Exemplo/exercício (iv)

Qual o IC a 95% de b? Intepre te.

```
# Manualmente
beta <- resumo_pai$coefficients[2, 1]
t_df <- c(qt(.025, df = mod_pai$df.residual),
          qt(.975, df = mod_pai$df.residual))
print(c(beta + t_df[1] * se_b,
        beta + t_df[2] * se_b))
```

```
## [1] 0.2949434 0.4740667
```

```
# Automatico
confint(mod_pai)
```

```
##                   2.5 %      97.5 %
## (Intercept) 86.2023282 117.7052895
## pai          0.2949434   0.4740667
```

Exemplo/exercício (v)

Faça o teste com $H_0 : \beta = 0$ e $H_a : \beta \neq 0$? Inteprete.

Exemplo/exercício (v)

Faça o teste com $H_0 : \beta = 0$ e $H_a : \beta \neq 0$? Inteprete.

```
print(resumo_pai$coefficients)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 101.953809 8.02618067 12.702656 3.270390e-34
## pai           0.384505 0.04563621  8.425437 1.349808e-16
```

MQO é BLUE se pressupostos valerem

Pressupostos

1. Linearidade
2. Exogeneidade estrita
3. Sem colinearidade perfeita
4. Amostragem aleatória
5. Homoscedasticidade

- Dado 1 & 2, ausência de viés.
    - Dado 3, modelo é estimável
    - Dado 1 a 5, ausência de viés + eficiência.

Estimativas de ponto

$$a = \bar{y} - b\bar{x} \qquad\qquad b = \frac{cov(x, y)}{var(x)}$$

No R, estima-se por `lm(y ~ x , data = xyz)` e `summary(obj)` em que `obj` é o objeto em que `lm()` foi gravado.

Estatísticas de ajuste

O coeficiente de determinação r^2 indica a proporção da variância de y "explicada" pelo modelo.

O estimador do erro padrão da regressão s_e quantifica o tamanho médio do resíduo.

No R, se o `summary()` for gravado em `resumo`, basta consultar `resumo$r.squared` e `resumo$sigma`.

Variância dos estimadores

Incerteza aumenta conforme s_e aumenta e diminui conforme n e $var(x)$ aumentam.

Distribuição amostral dos estimadores

Assintoticamente normal conforme a amostra cresce; para amostras pequenos é preciso o pressuposto de normalidade dos erros para que distribuição amostral seja normal.

ICs e testes de hipóteses

Construção muito parecida com o que já vimos, com se_b como estimador do erro padrão do coeficiente de x.

Cuidados!

1. Validade dos pressupostos
   - Inspeção visual da forma funcional é altamente recomendada
   - Muito cuidado com exogeneidade estrita antes de interpretar causalmente
2. Outliers

Séries temporais, painéis longitudinais etc violam esses pressupostos!

Próxima aula

Atividade

A atividade #6 será postada no Github no dia **28/11**, com entrega prevista para até **5/12**.

Leituras obrigatórias

Agresti 2018, cap. 11 a 13

Leituras optativas

Agresti 2018, cap. 11 a 13

Bussab e Morettin 2010 cap. 16

Huntington-Klein, cap. 13