Lição de Casa 2

James Hunter, Ph.D.

2 de outubro de 2020

Nesta lição da casa, vamos trabalhar com algumas problemas verdadeiras e alguns simplificados. As respostas precisam ser submetidas antes de **16 de outubro** por email: jameshunterbr@gmail.com.

Os dados ficam no arquivo "melanoma_raw.rds" no GitHub.

Resumo dos Dados

Dados são medidas feitos em pacientes com melanoma maligno. Cada paciente teve o tumor cirurgicamente removido no Departamento de Cirurgia Plástica no Hospital Universitário de Odense, Dinamarca durante 1962 ate 1977. Entre as medidas foram a espessura do tumor se uma úlcera for presente ou não. Os pesquisadores quiseram determinar se essas duas caraterísticas (espessura grande e presença de ulceração) aumentou a probabilidade de morte por causa de melanoma. Pacientes foram seguidos até o fim de 1977.

Fonte:

Angelo Canty and Brian Ripley (2019). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-23. Dados vêm de: Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993) **Statistical Models Based on Counting Processes**. Springer-Verlag.

Dicionário dos Dados do Arquivo `melanoma_raw.rds`

- `time` (tempo) (num)
  - Tempo de sobrevivência em dias desde a cirurgia, possivelmente censurado.
- `status` (estado) (num)
  - O estado dos pacientes ao final do estudo. "1" - morreu da melanoma "2" - vivo "3" - morreu de uma causa não relacionada à melanoma.
- `sex` (gênero) (num)
  - "1" - masculino
  - "2" - feminino
- `age` (idade) (num)
  - Idade em anos na data da cirurgia
- `year` (ano) (num)
  - Ano da cirurgia
- `thickness` (espessura) (num)
  - Espessura do tumor em mm
- `ulcer` (úlcera) (num)
  - "1" - úlcera presente

– "2" - úlcera ausente

Trabalho Preliminar

Antes de montar os gráficos, vai precisar fazer um pouco de limpeza de dados. Quais variáveis são realmente categóricas e devem estar traduzidas aos factors?

```
suppressPackageStartupMessages(library(tidyverse))
mel <- readRDS(here::here("melanoma_raw.rds")) %>%
  mutate(status = factor(status, labels = c("morreu_melanoma", "vivo", "morreu_outro")),
         sex = factor(sex, labels = c("masculino", "feminino")),
         ulcer = factor(ulcer, labels = c("presente", "ausente")))
glimpse(mel)
```

```
## Rows: 205
## Columns: 7
## $ time      <dbl> 10, 30, 35, 99, 185, 204, 210, 232, 232, 279, 295, 355, 3...
## $ status    <fct> morreu_outro, morreu_outro, vivo, morreu_outro, morreu_me...
## $ sex       <fct> feminino, feminino, feminino, masculino, feminino, femini...
## $ age       <dbl> 76, 56, 41, 71, 52, 28, 77, 60, 49, 68, 53, 64, 68, 63, 1...
## $ year      <dbl> 1972, 1968, 1977, 1968, 1965, 1971, 1972, 1974, 1968, 197...
## $ thickness <dbl> 6.76, 0.65, 1.34, 2.90, 12.08, 4.84, 5.16, 3.22, 12.88, 7...
## $ ulcer     <fct> ausente, presente, presente, presente, ausente, ausente, ...
```

Problema 1

Parte A - Gráficos

Gráfico 1

Qual é a distribuição das idades dos pacientes no estudo? Mostre através de uma histograma construída com `ggplot` e a `geom_histogram()`.

```
mel %>%
  ggplot(aes(x = age)) +
  geom_histogram(bins = 30)
```

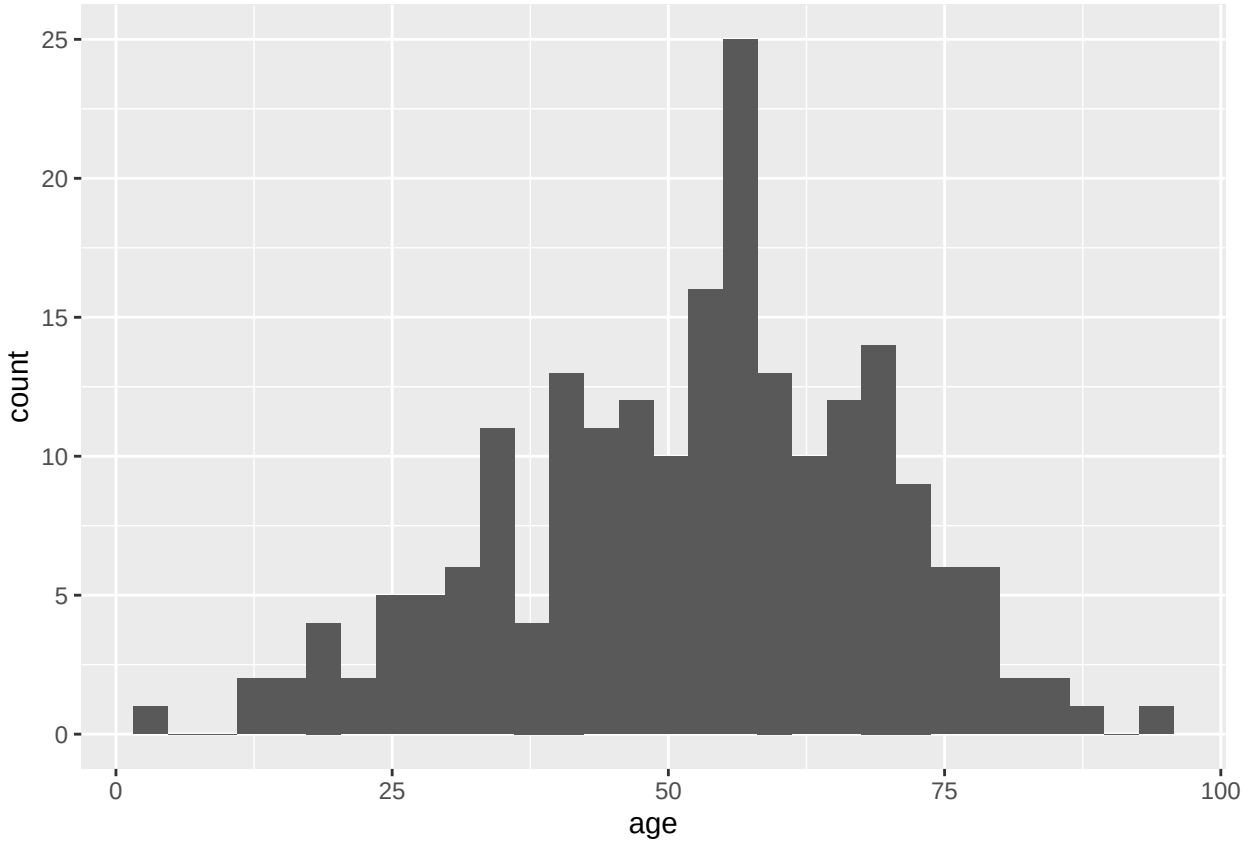


Gráfico 2

Qual é a relação entre tempo de sobrevivência (`tempo`) e idade? Existe uma associação? Mostre usando um scatterplot construído com `ggplot` e `geom_point()`.

```
ggplot(mel, aes(x = age, y = time)) +
  geom_point() +
  geom_smooth(method = "lm", color = "darkred")
```

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

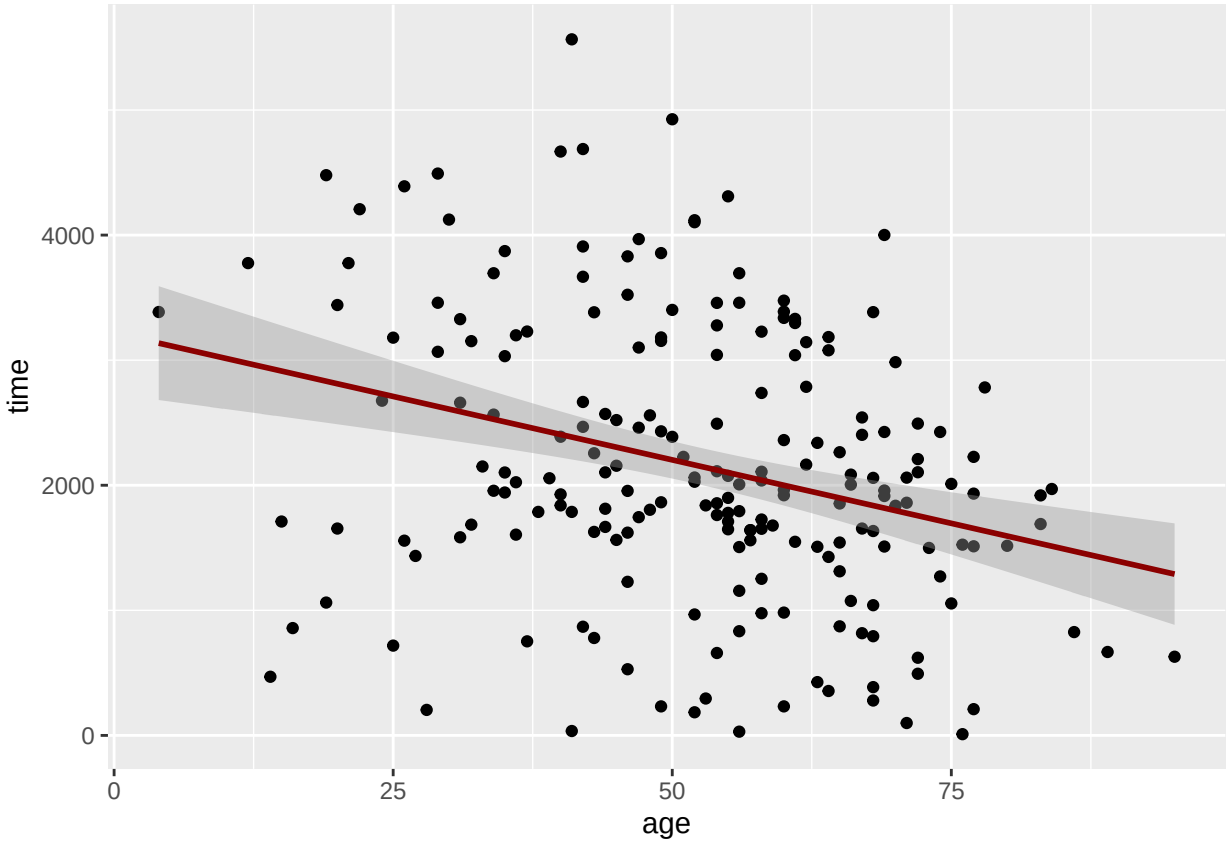


Gráfico 3

Existe uma diferença entre as espessuras dos tumores dos homens e mulheres? Faça um boxplot mostrando essa diferença, incluindo os pontos dos pacientes. Utilize `ggpubr` e sua função `ggboxplot`.

```
suppressPackageStartupMessages(library(ggpubr))
mel %>%
  ggboxplot(x = "sex",
            y = "thickness",
            palette = "uchicago",
            color = "sex",
            add = "jitter",
            theme = theme_gray())
```

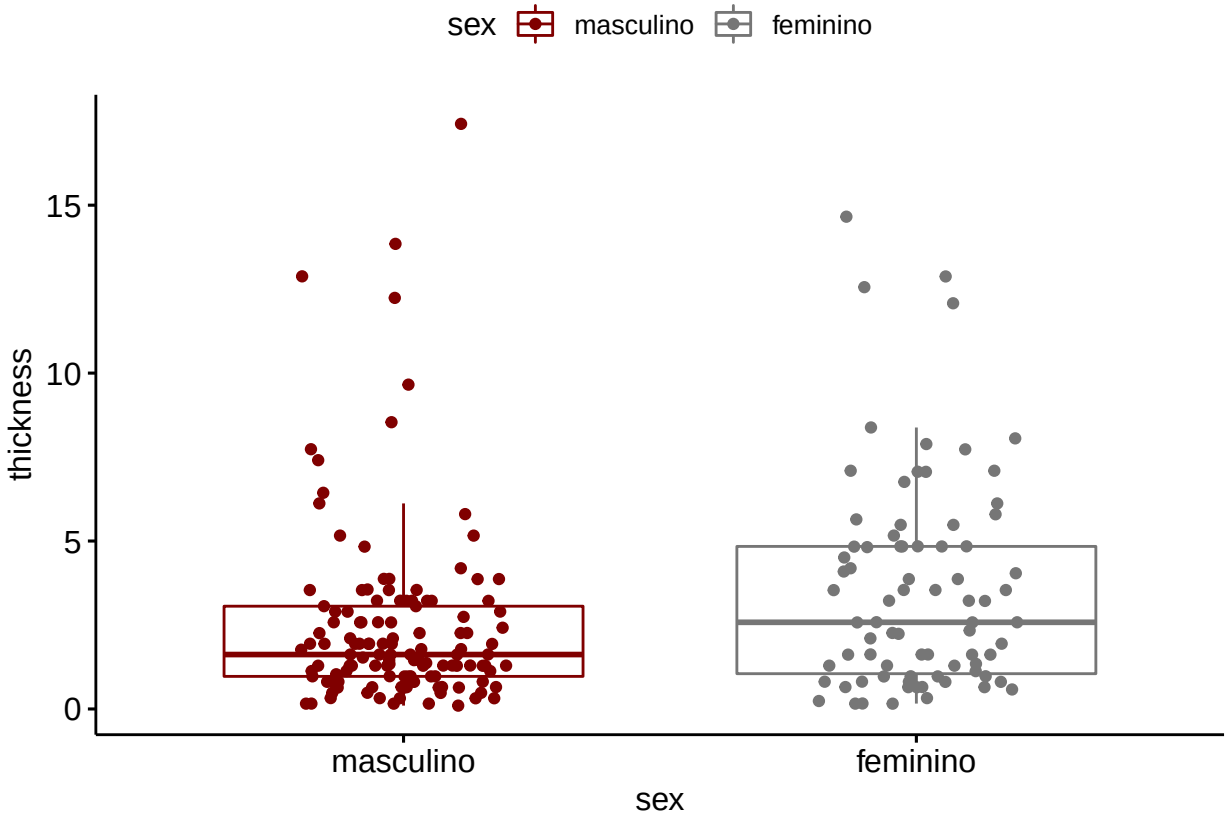
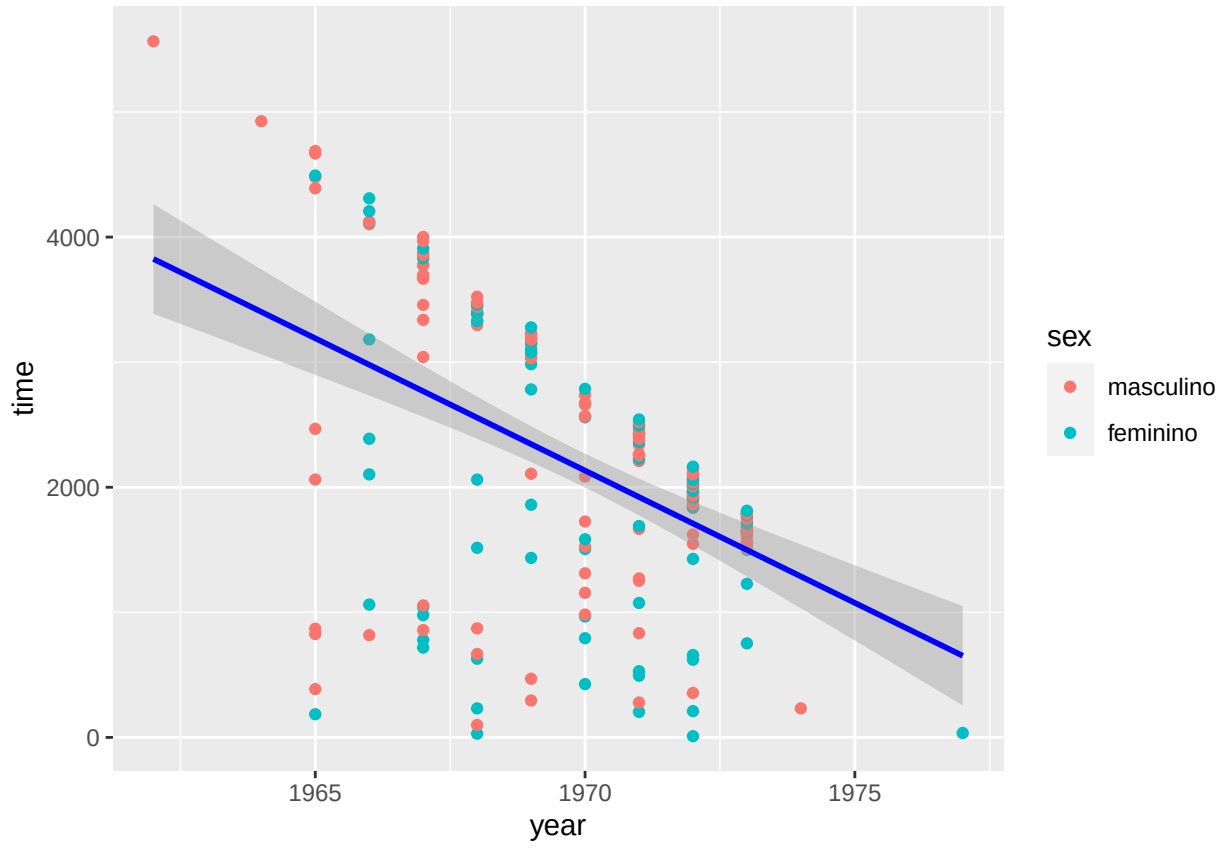


Gráfico 4

Existe differenças na sobrevivência das pessoas baseada no ano em que os pacientes fizeram a cirurgia? Mostre isso com scatterplot que inclui uma linha de tendência (`geom_smooth(method = "lm")`). Também, deve incluir o gênero do paciente como cor. Pode usar qualquer umas das funções que aprenderam. NB, este gráfico é mais complicado que os outros. Planeje ele bem antes de sentar em frente do RStudio.

```
mel %>%
  ggplot(aes(x = year, y = time, colour = sex)) +
  geom_point() +
  geom_smooth(method = "lm", colour = "blue")
```

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



Parte B - Estatística Descritiva

Use as funções `freq()` e `descr()` de `summarytools` para fazer um resumo dos variáveis de `melanoma`.

```
mel %>%
  select(status, sex, ulcer) %>%
summarytools::freq()
```

```
## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp
```

```
## Frequencies
## mel$status
## Type: Factor
##
##                       Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## -------------------- ------ --------- -------------- --------- --------------
##       morreu_melanoma     57     27.80          27.80     27.80          27.80
##                 vivo    134     65.37          93.17     65.37          93.17
##          morreu_outro     14      6.83         100.00      6.83         100.00
##                 <NA>      0                                0.00         100.00
##                Total    205    100.00         100.00    100.00         100.00
##
## mel$sex
## Type: Factor
##
```

```
##                   Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------- ------ --------- -------------- --------- --------------
##       masculino    126     61.46          61.46     61.46          61.46
##        feminino     79     38.54         100.00     38.54         100.00
##            <NA>      0                               0.00         100.00
##           Total    205    100.00         100.00    100.00         100.00
##
## mel$ulcer
## Type: Factor
##
##                  Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ---------------- ------ --------- -------------- --------- --------------
##       presente    115     56.10          56.10     56.10          56.10
##        ausente     90     43.90         100.00     43.90         100.00
##           <NA>      0                               0.00         100.00
##          Total    205    100.00         100.00    100.00         100.00
```

```
mel %>%
  select(time, age, year, thickness) %>%
  summarytools::descr(stats = "common")
```

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## Descriptive Statistics
## mel
## N: 205
##
##                     age   thickness      time      year
## ---------------- -------- ----------- --------- ---------
##             Mean    52.46        2.92   2152.80   1969.91
##          Std.Dev    16.67        2.96   1122.06      2.58
##              Min     4.00        0.10     10.00   1962.00
##           Median    54.00        1.94   2005.00   1970.00
##              Max    95.00       17.42   5565.00   1977.00
##          N.Valid   205.00      205.00    205.00    205.00
##        Pct.Valid   100.00      100.00    100.00    100.00
```

Problema 2

Existe uma diferença entre a sobrevivência das mulheres e homens depois da cirurgia? Também pode usar um t-test.

```r
t.test(mel$time[mel$sex == "masculino"], mel$time[mel$sex == "feminino"])
```

```
##
##  Welch Two Sample t-test
##
## data:  mel$time[mel$sex == "masculino"] and mel$time[mel$sex == "feminino"]
## t = 2.0848, df = 159.27, p-value = 0.03868
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   17.74767 656.12032
## sample estimates:
## mean of x mean of y
##  2282.643  1945.709
```

Problema 3

Reorganize a variável status para diferenciar entre óbito por melanoma e outros resultados. A pergunta é se mais pessoas têm óbito se tiverem tumores ulcerados.

```r
mel <- mel %>%
  mutate(obito_mel = ifelse(status == "morreu_melanoma", "mel", "outro"))
gmodels::CrossTable(x = mel$obito_mel, y = mel$ulcer, chisq = TRUE, format = "SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## | Chi-square contribution |
## |             Row Percent |
## |          Column Percent |
## |           Total Percent |
## |-------------------------|
##
## Total Observations in Table:  205
##
##                | mel$ulcer
## mel$obito_mel  | presente  |  ausente  | Row Total |
## ---------------|-----------|-----------|-----------|
##            mel |       16  |       41  |       57  |
##                |    7.982  |   10.199  |           |
##                |   28.070% |   71.930% |   27.805% |
##                |   13.913% |   45.556% |           |
##                |    7.805% |   20.000% |           |
## ---------------|-----------|-----------|-----------|
##          outro |       99  |       49  |      148  |
##                |    3.074  |    3.928  |           |
##                |   66.892% |   33.108% |   72.195% |
##                |   86.087% |   54.444% |           |
##                |   48.293% |   23.902% |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |      115  |       90  |      205  |
##                |   56.098% |   43.902% |           |
## ---------------|-----------|-----------|-----------|
##
```

```
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  25.18254     d.f. =  1     p =  5.215219e-07
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  23.6309     d.f. =  1     p =  1.166987e-06
##
##
##        Minimum expected frequency: 25.02439
```

[Answer: We reject the Null that deaths occur independent of ulcerated tumors. In fact, it appears that of the 57 people who died as a result of melanoma, only 28.07% had ulcerated tumors. Also of the 115 people with ulcerated tumors only 13.91% died as a result of melanoma. The others have remained alive or died of other causes.]

Problema 4

Conduzir uma regressão linear utilizando `lm()` da relação possível entre idade e tempo de sobrevivência. Mostrar o resultado com `summary()`. Extrair os coeficientes com `broom::tidy()`.

```
fit <- lm(time ~ age, data = mel)
summary(fit)
```

```
##
## Call:
## lm(formula = time ~ age, data = mel)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2464.3  -646.2   -54.4   712.1  3179.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3217.448    247.879  12.980  < 2e-16 ***
## age          -20.293      4.504  -4.506 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1072 on 203 degrees of freedom
## Multiple R-squared:  0.09091,    Adjusted R-squared:  0.08643
## F-statistic:  20.3 on 1 and 203 DF,  p-value: 1.116e-05
```

```
broom::tidy(fit)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   3217.     248.       13.0  1.90e-28
## 2 age            -20.3      4.50     -4.51 1.12e- 5
```