Stat 140: Inference for Simple Linear Regression Example - Cognitive Decline

Evan Ray

December 4, 2017

Age and Cognitive Function

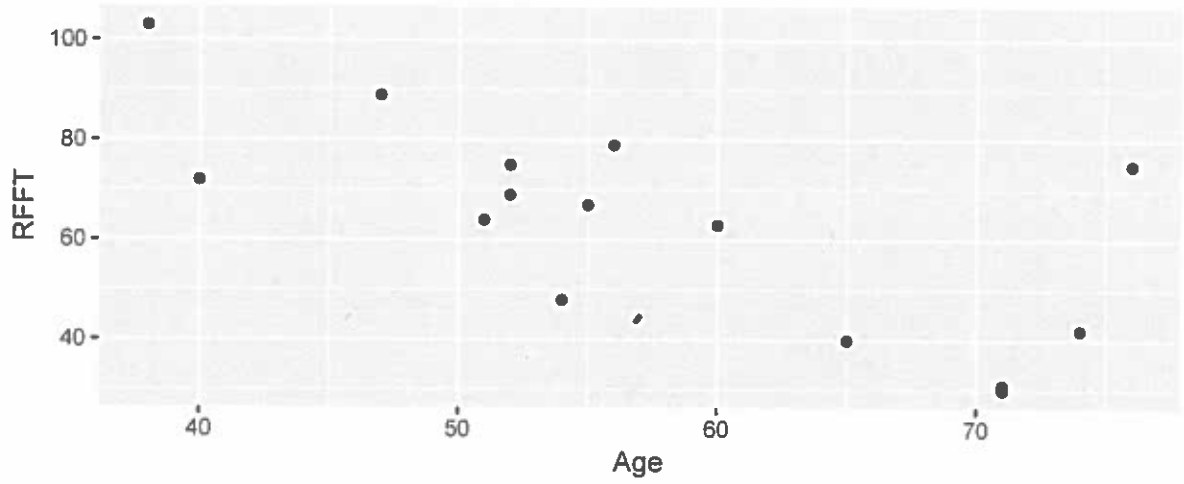
Various demographic and cardiovascular risk factors, including assessments of cognitive function, were collected as a part of the Prevention of Renal and Vascular End-stage Disease study, which took place between 2003 and 2006 in the Netherlands. Cognitive function was measured with the The Ruff Figural Fluency Test (RFFT). The test consists of drawing as many unique designs as possible from a pattern of dots under timed conditions; scores range from 0 to 175 points (worst and best score, respectively).

Let's examine the relationship between the RFFT score and age. Here is a plot of the data for a random sample of 15 individuals in the study, as well as results of a linear model fit to the data.

```
cognitive_decline <- read_csv("https://mhc-stat140-2017.github.io/data/openintro/statins/statins.csv")
```

```
set.seed(123)
cognitive_decline_sample <- sample_n(cognitive_decline, size = 15)
```

```
ggplot() +
  geom_point(mapping = aes(x = Age, y = RFFT), data = cognitive_decline_sample)
```




```
lm_fit_sample <- lm(RFFT ~ Age, data = cognitive_decline_sample)
```

```
summary(lm_fit_sample)

##
## Call:
## lm(formula = RFFT ~ Age, data = cognitive_decline_sample)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19.50 -13.39  -0.30   8.83  35.22
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  135.546     20.583    6.59  1.8e-05 ***
## Age           -1.260      0.351   -3.59   0.0033 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.6 on 13 degrees of freedom
## Multiple R-squared:  0.497,  Adjusted R-squared:  0.459
## F-statistic: 12.9 on 1 and 13 DF,  p-value: 0.00331
```

1. Hypothesis Tests

(a) Write down the null and alternative hypotheses for a test that the slope of a line describing the relationship between age and RFFT scores in the population is 0.

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

(b) Draw a picture of a relevant t distribution and shade in the area corresponding to the p-value for this test.

The test statistic is $\frac{b_1 - \beta_1}{SE(b_1)} = \frac{-1.26 - 0}{0.351} = -3.59$

If the null hypothesis is true, this follows a t_{n-2} distribution

The p-value is the probability of getting a test statistic at least as extreme as this assuming the null hypothesis is true. That's the shaded area in the picture to the left. "At least as extreme" means to the left or to the right, because the alternative hypothesis is $H_A: \beta_1 \neq 0$ which doesn't specify a direction.

(c) Use the following R output to calculate the p-value for the test (your result should agree with the p-value in the R output above). Make sure you understand why this command is the right one to use for this purpose.

```
pt(-3.587, df = 13)

## [1] 0.001657
```

This is the "lower half" of the p-value, i.e. the probability of getting a test statistic less than -3.59.

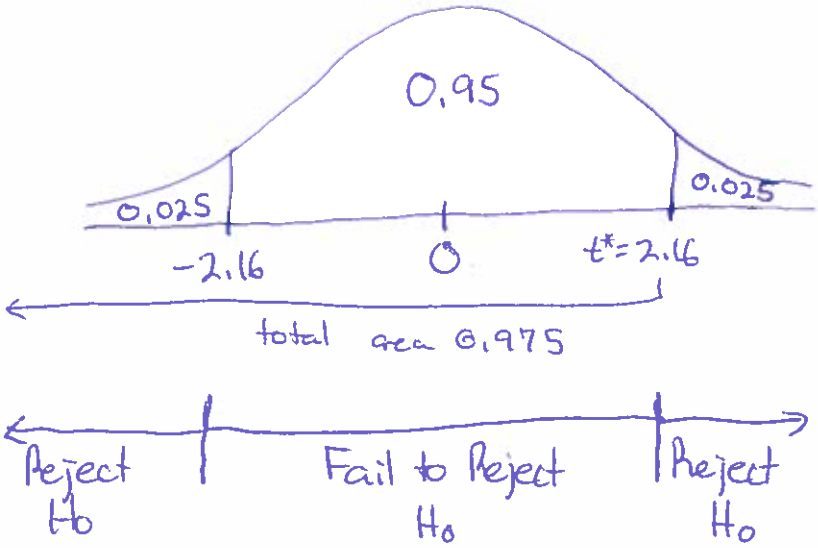
The p-value $= 2 * 0.001657 \approx 0.0033$

This matches the R output above.

(d) We will do this as a class. Conduct the hypothesis test at the $\alpha = 0.05$ level by comparing the test statistic to a critical value. Use the following output from R:

```
qt(0.975, df = 13)

## [1] 2.16
```



The test statistic is -3.59, which is farther away from 0 than the critical value of $t^* = 2.16$

This means that we can reject the null hypothesis.

The data offer enough evidence to conclude that the slope in the population is different from 0, at the $\alpha = 0.05$ significance level.

2. Confidence Intervals

(a) Use the following R output, as well as information from the linear model summary above, to calculate a 95 percent confidence interval for the slope of a line describing the relationship between age and RFFT scores in the population. Make sure you understand why this command is the right one to use for this purpose.

```
qt(0.975, df = 13)
```

```
## [1] 2.16
```

Our confidence interval will be of the form $b_1 \pm t^* \cdot SE(b_1)$, where $b_1 = -1.26$ (R output on p. 2), $SE(b_1) = 0.351$ (R output on p. 2), and $t^* = 2.16$

$-1.26 - 2.16 * 0.351 = -2.018$, and $-26 + 2.16 * 0.351 = -0.502$.

We are 95% confident that the slope of the regression line describing the relationship between age and RFFT scores is between -2.018 and -0.502. If we took many different samples and calculated a 95% C.I. based on the data in each sample, about 95% of those confidence intervals would contain the true population slope β_1.

(b) Here is the output from a linear model fit using the full data set of all 4095 people in the study. Use this output to obtain a 95 percent confidence interval for the slope of a line describing the relationship between age and RFFT scores in the population.

```
lm_fit <- lm(RFFT ~ Age, data = cognitive_decline)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = RFFT ~ Age, data = cognitive_decline)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68.37 -15.64  -1.06  14.80  79.28
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   132.34       1.68    79.0   <2e-16 ***
## Age            -1.17       0.03   -38.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.3 on 4093 degrees of freedom
## Multiple R-squared:  0.27,   Adjusted R-squared:  0.269
## F-statistic: 1.51e+03 on 1 and 4093 DF,  p-value: <2e-16
```

We also need to know that $qt(0.975, df = 4093) = 1.96$.

$b_1 \pm t^* \cdot SE(b_1)$

$-1.17 \pm 1.96 \cdot 0.03$

$[-1.23, -1.11]$

We are 95% confident that the population slope is between -1.23 and -1.11, etc. etc.

(c) Compare the confidence intervals you got in parts (a) and (b). Which is wider?

The interval with a larger sample size is narrower.

The main reason for this is that the ~~sample~~ standard error of b_1 is smaller with a larger sample size.

The formula for $SE(b_1)$ is:

$$SE(b_1) = \frac{s_e}{\sqrt{n-1}\, s_x}$$

s_e ← residual standard deviation

s_x ← standard deviation of x values

n ← sample size

As the sample size increases, we are dividing by a larger value so $SE(b_1)$ decreases.

We have seen this basic idea before with confidence intervals for the population proportion and the population mean. In all of those cases, confidence intervals tend to get narrower as the sample size increases.