Social Statistics

Introducing Regression

November 15, 2021

Quick PS2 Shortcuts

1. Without using any R shortcuts, find the 95% confidence interval for the mean of `eqwlth` in 2010, 2014, and 2018.

```
q1 <- ps2 |>
  filter(year == 2010 | year == 2014 | year == 2018) |>
  group_by(year) |>
  summarize(mean = mean(eqwlth),
            sd = sd(eqwlth),
            n = length(eqwlth),
            se = sd / sqrt(n),
            ll = mean - 1.96*se,
            ul = mean + 1.96*se)
```

Quick PS2 Shortcuts

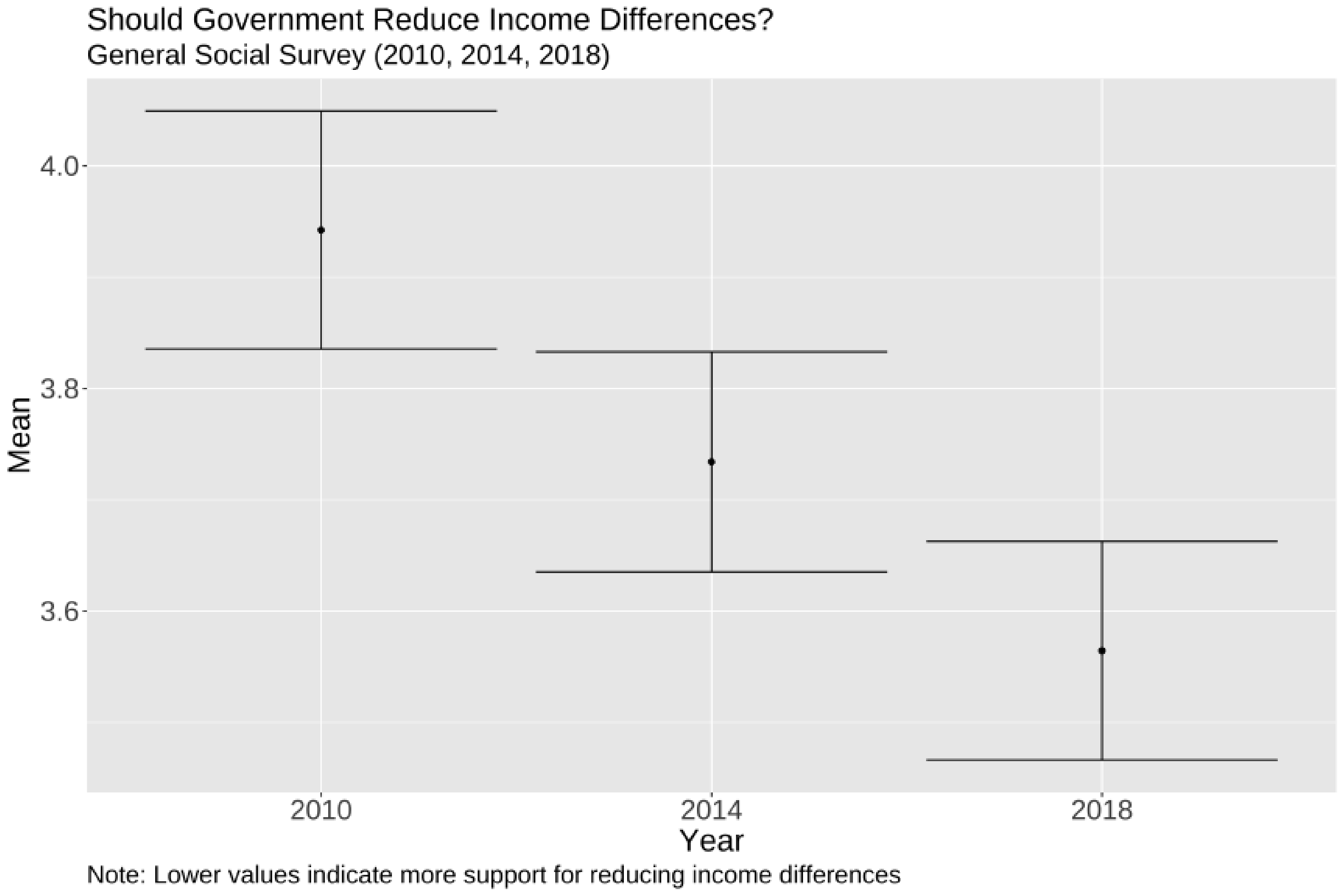
year	mean	sd	n	se	ll	ul
2010	3.942	2.008	1355	0.055	3.835	4.049
2014	3.734	2.057	1666	0.050	3.635	3.833
2018	3.564	1.960	1529	0.050	3.466	3.663

Quick PS2 Shortcuts

```
plot_q1 <- ggplot(q1, aes(x = year, y = mean,
                          ymin = ll, ymax = ul))

plot_q1 + geom_point() + geom_errorbar() +
labs(x = "Year",
     y = "Mean",
     title = "Should Government Reduce Income Differences?",
     subtitle = "General Social Survey (2010, 2014, 2018)",
     caption = "Note: Lower values indicate more support for reducing income
     theme(plot.caption = element_text(hjust = 0))
```

Quick PS2 Shortcuts



Quick PS2 Shortcuts

2. Which (if any) age categories showed significant differences in mean `eqwlth` scores between the 2010 and 2018 surveys?

```
multiple_ttests <- ps2 |>
  filter(!is.na(agecat)) |>
  group_by(agecat) |>
  summarise(across(eqwlth,
                   list( # To capture multiple values from tests
                     (~t.test(.[year == 2010],
                            .[year == 2018])$statistic),
                    ~t.test(.[year == 2010],
                            .[year == 2018])$p.value
                    ) # Close list
                   ) # Close across
             ) # Close summarise

colnames(multiple_ttests) <- c("Age Category", "Test Statistic",
                               "P Value")
```

Quick PS2 Shortcuts

Age Category	Test Statistic	P Value
1	1.267	0.206
2	2.760	0.006
3	2.928	0.004
4	1.414	0.158
5	2.792	0.005

Where We've Been

Descriptive statistics gave us means, standard deviations

- "What are the spreads and the shapes of our observed distributions?"

Probability gave us ways to use our sample statistics to predict ranges of possible population parameters

- "What is the likelihood of getting the values we observe?"

Inference gave us tools to test significance

- "What is the likelihood of getting a value more extreme than the values we observe?"

Two Things We Still Want

1. Better conclusions

- Asssociations peaked with correlation
- If correlation coefficient tells us that X and Y *tend to move together*, regression tells us *how much* they tend to move together

2. Explanations of variation

- Inference offered us ways to know if X and Y are dependent or independent (Chi-squared Test, Fisher's Test, etc.)
- Dependent associations may be influenced by *confounding*.

Start With Regression Basics

Basic assumption (for now): The relationship between X and Y is linear

- HS Flashback: y = mx + b, where m is the slope and b is the intercept

Linear relationship is regression equation:

$$\hat{y}_i = \alpha + \beta X_i + \epsilon_i$$

- Read as: *regress y on x*

Start With Regression Basics

$$\hat{y}_i = \alpha + \beta X_i + \epsilon_i$$

- $\hat{y}_i$ = predicted outcome, the best guess
- α = intercept or constant, where the line hits the y-axis when x is 0
- β = the slope, the multiplier for every X, known as the coefficient
- X_i = the observed value of X
- ϵ_i = error (or residual), difference between observed and predicted values

Example from UN Human Development Project

Example - Schooling & Life Expectancy

Fitting The Regression Line

Recall that a *residual* is the difference between the observed value, y, and the predicted value on the line, $\hat{y}$

We want a line that makes every residual as small as possible

Every observation has a residual. How do we combine them?

- Can't just add them up since negatives could cancel out positives
- Absolute values are the usual fix, but they don't help as much this time since they offer little guide for where to start with α and β

Fitting The Regression Line

Sum of the squared residuals gets us closest

- $SSE = \sum (y - \hat{y})^2$
- Line with the smallest sum has the *least squares*: why basic regression is called *Ordinary Least Squares*

Squaring gives extra weight to biggest residuals (the observations that a given line does a particularly bad job at including)

To find beta and alpha, we'll use basics we have seen: how the observed x's differ from the mean of x, how the observed y's differ from the mean of y, and how the distribution of x and y tend to move together

Fitting Beta and Alpha

Let's try the example of regressing life expectancy in years on the standardized schooling expectancy

Start with basic descriptives

- What's the correlation between the two variables?
- What are the mean and standard deviation of `std_schooling_expectancy`?
- What are the mean and standard deviation of `life_expectancy`?

Finding Beta and Alpha

```
# Correlation
cor(hdi$std_schooling_expectancy, hdi$life_expectancy)
```

```
## [1] 0.8061841
```

Interpretation?

Finding Beta and Alpha

```
# Mean and Standard Deviation of X
mean(hdi$std_schooling_expectancy)
```

```
## [1] -5.031447e-11
```

```
sd(hdi$std_schooling_expectancy)
```

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

```
# Mean and Standard Deviation of Y
mean(hdi$life_expectancy)
```

```
## [1] 71.83705
```

```
sd(hdi$life_expectancy)
```

```
## [1] 8.165182
```

Fitting The Regression Line

We have all we need to find beta:

$$\beta = cor_{xy}\frac{s_y}{s_x}$$

And beta will be the missing piece to help us find alpha:

$$\alpha = \bar{y} - \beta\bar{x}$$

Finding Beta

$$\beta = cor_{xy}\frac{s_y}{s_x}$$

```
beta <- cor(hdi$std_schooling_expectancy,
      hdi$life_expectancy) *
     (sd(hdi$life_expectancy) /
      sd(hdi$std_schooling_expectancy))

beta
```

```
## [1] 6.58264
```

Interpreting Beta

Every one unit increase in the value of X is associated with an increase of beta in the predicted value of Y, on average

- In this model, a one standard deviation increase in schooling expectancy is associated with an increase of 6.5826 years in life expectancy, on average

And since we are working with linear regression, a one unit decrease in the value of X is associated with a decrease of beta in the predicted value of Y, on average

- In this model, a one standard deviation decrease in schooling expectancy is associated with a decrease of 6.5826 years in life expectancy, on average

Finding Alpha

$$\alpha = \bar{y} - \beta\bar{x}$$

```
alpha <- mean(hdi$life_expectancy) -
      beta*(mean(hdi$std_schooling_expectancy))

alpha
```

```
## [1] 71.83705
```

When X is 0, our model predicts that Y should be 71.8371

In this case (since x is standardized with a mean of 0), a country with a schooling expectancy at the average of the distribution would be predicted to have a life expectancy of 71.8371 years.

Fitting The Regression Line

Now we have our line: y = 71.8371 + 6.5826X

Let's add it to our plot using `geom_abline()`:

```
schooling_life_plot1 <- ggplot(hdi, aes(
     x = std_schooling_expectancy, y = life_expectancy))

schooling_life_plot1 + geom_point(color = "Dark Gray") +
     labs(x = "Standardized Schooling Expectancy",
      y = "Life Expectancy",
      title = "Schooling and Life Expectancies",
      subtitle = "(UNHDP, 2016)") +
     geom_abline(intercept = 71.8371, slope = 6.5826)
```

Fitting The Regression Line

Predicting Values of Y

If the line is correct, there should be a point on the line where X = 0 and Y = 71.8371

```
schooling_life_plot1 + geom_point(color = "Dark Gray") +
     labs(x = "Standardized Schooling Expectancy",
      y = "Life Expectancy",
      title = "Schooling and Life Expectancies",
      subtitle = "(UNHDP, 2016)") +
     geom_abline(intercept = 71.8371, slope = 6.5826) +
     geom_point(x = 0, y = 71.8371, color = "Red", size = 3)
```

Predicting Values of Y

Predicting Values of Y

Digging Deeper: when x increases by 1, $\hat{y}$ is expected to increase by 6.5826

So if x is 1 standard deviation above the mean, what is $\hat{y}$?
And if x is 1 standard deviation below the mean, what is $\hat{y}$?

Prediction always has to start with value of α!

```
predicted_y_plus1sd <- alpha + beta*1
predicted_y_plus1sd
```

```
## [1] 78.41969
```

```
predicted_y_minus1sd <- alpha + beta*-1
predicted_y_minus1sd
```

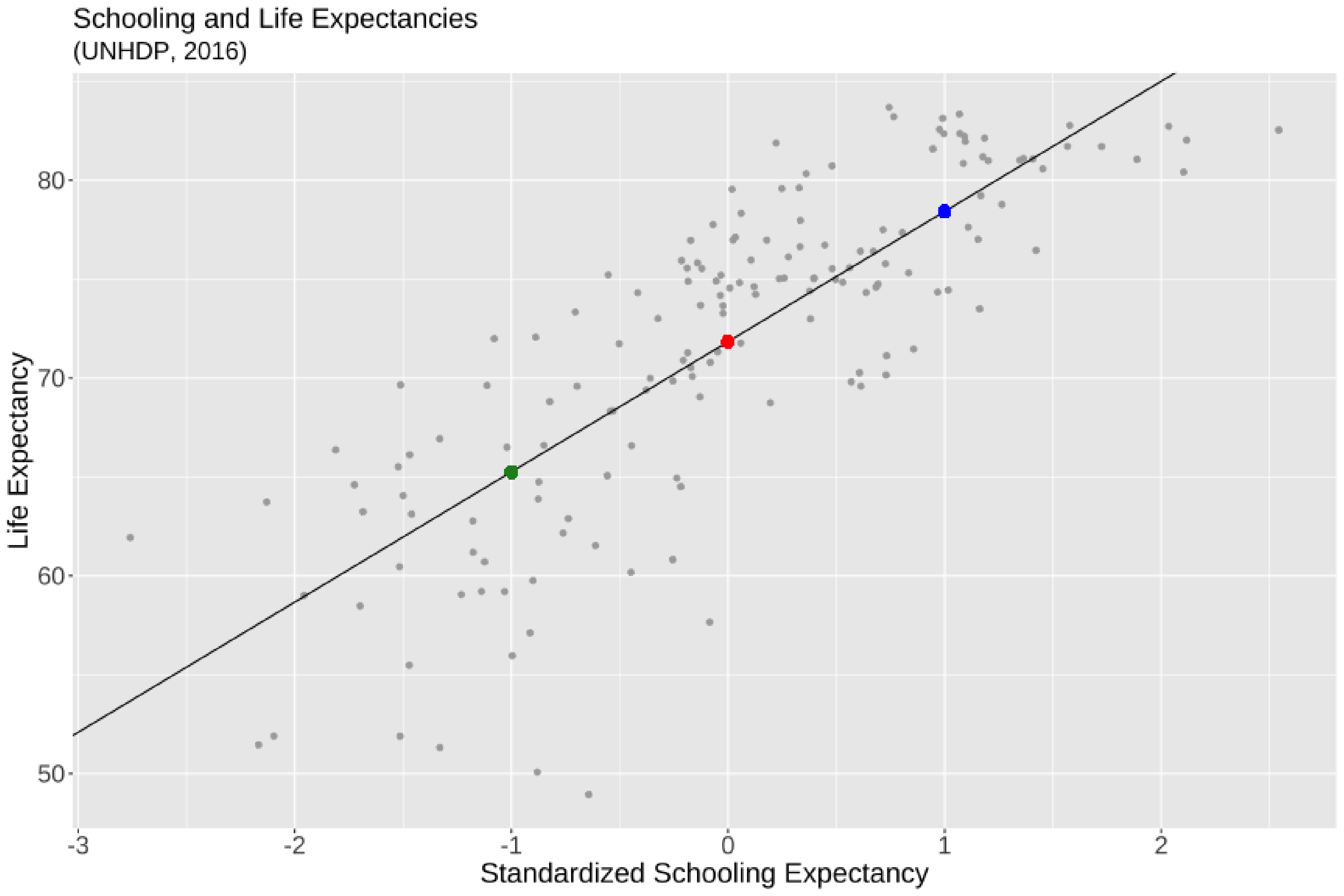
```
## [1] 65.25441
```

Predicting Values of Y

Put these points on our plot...

```
schooling_life_plot1 + geom_point(color = "Dark Gray") +
labs(x = "Standardized Schooling Expectancy",
y = "Life Expectancy",
      title = "Schooling and Life Expectancies",
      subtitle = "(UNHDP, 2016)") +
geom_abline(intercept = 71.8371, slope = 6.5826) +
geom_point(x = 0, y = 71.8371, color = "Red", size = 3) +
geom_point(x = 1, y = 78.4197, color = "Blue", size = 3) +
geom_point(x = -1, y = 65.2545, color = "Forest Green",
  size = 3)
```

Predicting Values of Y



Regression in R

As always, R makes this easier. Meet the `lm()` command.

```
# Start by saving the model as an object:

schooling_life_model1 <-
      lm(life_expectancy ~ std_schooling_expectancy,
        data = hdi)
```

```
# Then look at the summary of the saved model:

summary(schooling_life_model1)
```

Regression in R

Should look familiar: standard errors, t-stats, p-values!

```
> summary(schooling_life_model1)

Call:
lm(formula = life_expectancy ~ std_schooling_expectancy, data = hdi)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.6597  -2.4645   0.3544   3.4981   8.5817 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               71.8371     0.3843  186.91   <2e-16 ***
std_schooling_expectancy   6.5826     0.3856   17.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.846 on 157 degrees of freedom
Multiple R-squared:  0.6499,	Adjusted R-squared:  0.6477 
F-statistic: 291.5 on 1 and 157 DF,  p-value: < 2.2e-16
```

Regression in R

Red Box = Alpha; Blue Box = Beta

```
> summary(schooling_life_model1)

Call:
lm(formula = life_expectancy ~ std_schooling_expectancy, data = hdi)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.6597  -2.4645   0.3544   3.4981   8.5817 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               71.8371     0.3843  186.91   <2e-16 ***
std_schooling_expectancy   6.5826     0.3856   17.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.846 on 157 degrees of freedom
Multiple R-squared:  0.6499,	Adjusted R-squared:  0.6477 
F-statistic: 291.5 on 1 and 157 DF,  p-value: < 2.2e-16
```

R's Regression Output - Std Error

Std. Error = SE of the coefficient

```
> summary(schooling_life_model1)

Call:
lm(formula = life_expectancy ~ std_schooling_expectancy, data = hdi)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.6597  -2.4645   0.3544   3.4981   8.5817 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               71.8371     0.3843  186.91   <2e-16 ***
std_schooling_expectancy   6.5826     0.3856   17.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.846 on 157 degrees of freedom
Multiple R-squared:  0.6499,	Adjusted R-squared:  0.6477 
F-statistic: 291.5 on 1 and 157 DF,  p-value: < 2.2e-16
```

R's Regression Output - Std Error

$$se = \frac{s}{\sqrt{\sum{(x - \bar{x})^2}}}$$

and

$$s = \sqrt{\frac{\sum{(y - \hat{y})^2}}{n - 2}}$$

R's Regression Output - Std Error

The standard error formula uses the predicted values of y to calculate the residuals

R makes it easy to save all the predicted values from a model:

```
hdi$predicted_life_expectancy <-
    schooling_life_model1$fitted.values
```

R's Regression Output - Std Error

Now you can plug in the predicted values to the rest of the standard error equation:

```
se_numerator <- sqrt(sum((hdi$life_expectancy -
  hdi$predicted_life_expectancy)^2) /
    (length(hdi$life_expectancy) - 2))

se_denominator <- sqrt(sum((hdi$std_schooling_expectancy -
  mean(hdi$std_schooling_expectancy))^2))

se <- se_numerator / se_denominator

se
```

```
## [1] 0.3855599
```

R's Regression Output - Std Error

```
> summary(schooling_life_model1)

Call:
lm(formula = life_expectancy ~ std_schooling_expectancy, data = hdi)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.6597  -2.4645   0.3544   3.4981   8.5817 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               71.8371     0.3843  186.91   <2e-16 ***
std_schooling_expectancy   6.5826     0.3856   17.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.846 on 157 degrees of freedom
Multiple R-squared:  0.6499,	Adjusted R-squared:  0.6477 
F-statistic: 291.5 on 1 and 157 DF,  p-value: < 2.2e-16
```

R's Regression Output - T Value

t = test statistic for a t-test that coefficient differs from zero

```
> summary(schooling_life_model1)

Call:
lm(formula = life_expectancy ~ std_schooling_expectancy, data = hdi)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.6597  -2.4645   0.3544   3.4981   8.5817 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               71.8371     0.3843  186.91   <2e-16 ***
std_schooling_expectancy   6.5826     0.3856   17.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.846 on 157 degrees of freedom
Multiple R-squared:  0.6499,	Adjusted R-squared:  0.6477 
F-statistic: 291.5 on 1 and 157 DF,  p-value: < 2.2e-16
```

R's Regression Output - T Value

t = coefficient estimate / standard error

```
6.5826 / .3856
```

```
## [1] 17.07106
```

R's Regression Output - T Value

```
> summary(schooling_life_model1)

Call:
lm(formula = life_expectancy ~ std_schooling_expectancy, data = hdi)

Residuals:
     Min       1Q   Median       3Q      Max
-18.6597  -2.4645   0.3544   3.4981   8.5817

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               71.8371     0.3843  186.91   <2e-16 ***
std_schooling_expectancy   6.5826     0.3856   17.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.846 on 157 degrees of freedom
Multiple R-squared:  0.6499,	Adjusted R-squared:  0.6477
F-statistic: 291.5 on 1 and 157 DF,  p-value: < 2.2e-16
```

R's Regression Output - P Value

P>|t| = p-value for two-tailed test

```
> summary(schooling_life_model1)

Call:
lm(formula = life_expectancy ~ std_schooling_expectancy, data = hdi)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.6597  -2.4645   0.3544   3.4981   8.5817 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               71.8371     0.3843  186.91   <2e-16 ***
std_schooling_expectancy   6.5826     0.3856   17.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.846 on 157 degrees of freedom
Multiple R-squared:  0.6499,	Adjusted R-squared:  0.6477 
F-statistic: 291.5 on 1 and 157 DF,  p-value: < 2.2e-16
```

R's Regression Output - P Value

```r
# Area in right tail:
pr_tail <- 1 - pt(17.07, df = 157)

# Area in both tails (what output gives):
2 * pr_tail
```

```
## [1] 0
```

Can we reject the null hypothesis that the coefficient for `std_schooling_expectancy` is different from 0?

Yes, because `Pr(>|t|)` is less than .05

Note the stars!

R's Regression Output - P Value

```
> summary(schooling_life_model1)

Call:
lm(formula = life_expectancy ~ std_schooling_expectancy, data = hdi)

Residuals:
     Min       1Q   Median       3Q      Max
-18.6597  -2.4645   0.3544   3.4981   8.5817

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               71.8371     0.3843  186.91   <2e-16 ***
std_schooling_expectancy   6.5826     0.3856   17.07   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.846 on 157 degrees of freedom
Multiple R-squared:  0.6499,	Adjusted R-squared:  0.6477
F-statistic: 291.5 on 1 and 157 DF,  p-value: < 2.2e-16
```

Plotting Regressions

More common to use `geom_smooth(method = lm)` than `geom_abline()`:

```
schooling_life_plot1 + geom_point(color = "Dark Gray") +
     labs(x = "Standardized Schooling Expectancy",
          y = "Life Expectancy",
      title = "Schooling and Life Expectancies",
      subtitle = "(UNHDP, 2016)") +
      geom_smooth(method = lm)
```

Plotting Regressions

Exercise 1

Regress the gender inequality index on the percentage of members of parliament who are female

```
inequality_parliament_model <-
      lm(gender_inequality_index ~ female_parliament_pct,
            data = hdi)
```

```
summary(inequality_parliament_model)
```

Exercise 1

```
##
## Call:
## lm(formula = gender_inequality_index ~ female_parliament_pct,
##     data = hdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.165 -16.654  -0.566  14.986  34.203
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            48.1833     2.9745  16.199  < 2e-16 ***
## female_parliament_pct  -0.5728     0.1228  -4.665 6.56e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.77 on 157 degrees of freedom
## Multiple R-squared:  0.1217,    Adjusted R-squared:  0.1161
## F-statistic: 21.76 on 1 and 157 DF,  p-value: 6.563e-06
```

Exercise 1

Gender Inequality Index =

48.18 + (-0.5728 $\times$ Female Parliament Pct)

An increase of one point in the percentage of parliament members who are women is associated with a decrease in the gender inequality index of .573, on average.

In the US, the percentage of parliament members who are female is 19.48. What is the US' predicted value on the gender inequality index?

```
48.18 + (-.5728*19.48)
```

```
## [1] 37.02186
```

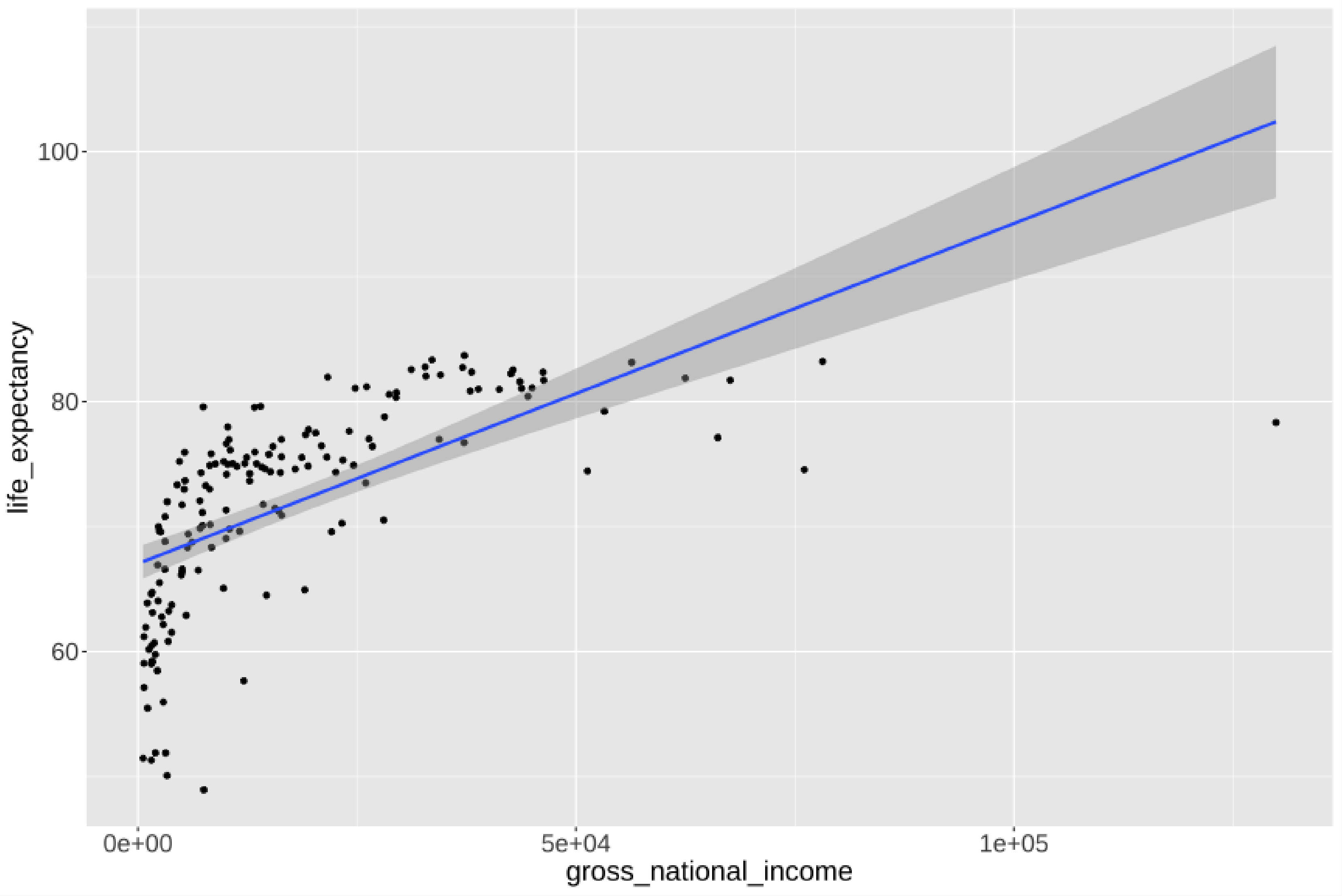
Exercise 2

What would you expect about the relationship between `gross_national_income` and `life_expectancy`?

```
income_life_expectancy_plot <- ggplot(hdi, aes(x = gross_national_income,
                                  y = life_expectancy)) + geom_point() +
  geom_smooth(method = lm)

income_life_expectancy_plot
```

Exercise 2



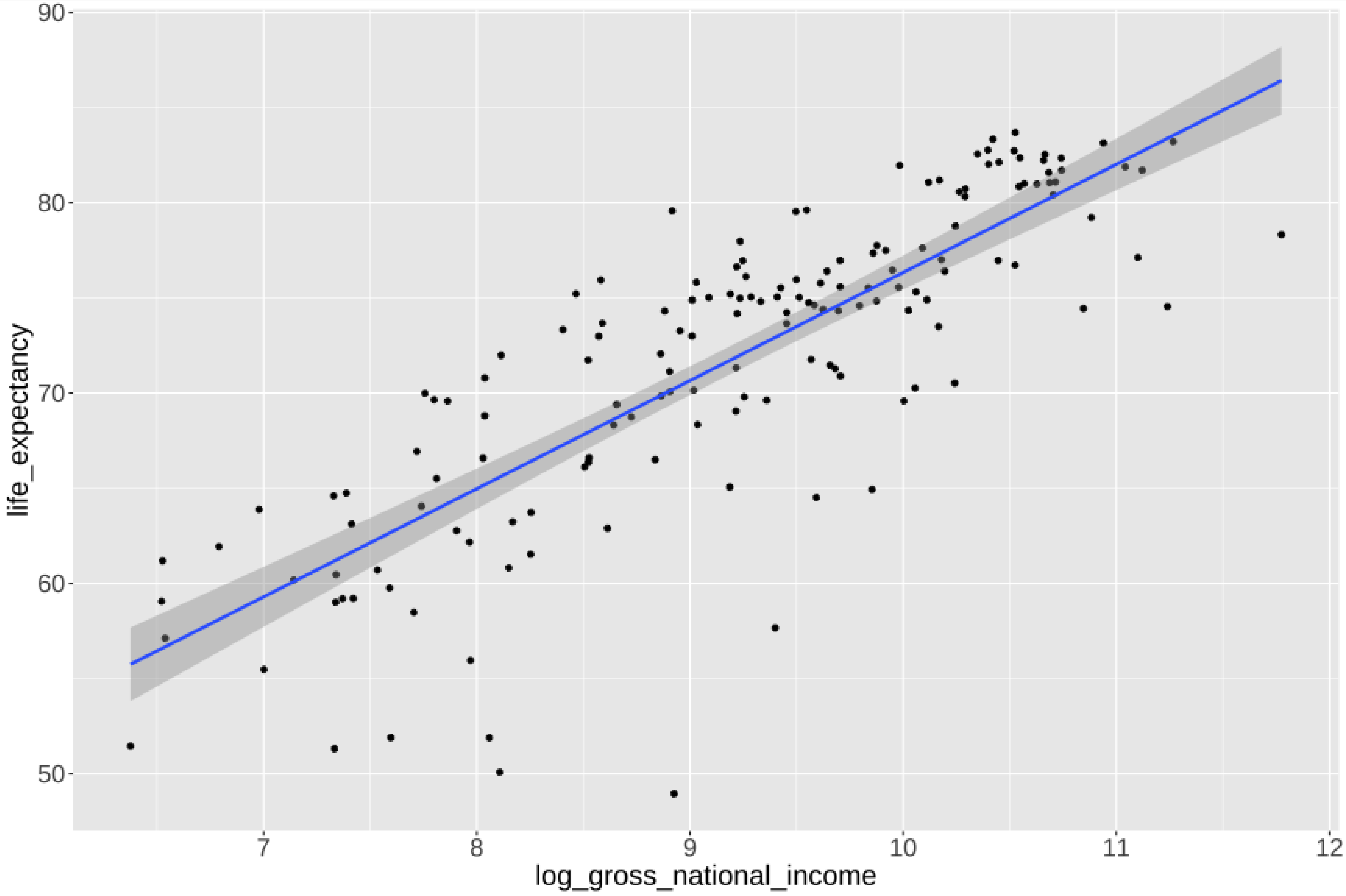
100
80
60
life_expectancy
0e+00
5e+04
1e+05
gross_national_income

Exercise 2

```
income_life_expectancy_plot <- ggplot(hdi,
                                      aes(x = log_gross_national_income,
                                       y = life_expectancy)) + geom_point() +
  geom_smooth(method = lm)

income_life_expectancy_plot
```

Exercise 2



life_expectancy
log_gross_national_income
90
80
70
60
50
7
8
9
10
11
12

Exercise 2

Try the regression model using `life_expectancy` and `log_gross_national_income`…

```
income_life_expectancy_model <-
      lm(life_expectancy ~ log_gross_national_income,
            data = hdi)

summary(income_life_expectancy_model)
```

Exercise 2

```
##
## Call:
## lm(formula = life_expectancy ~ log_gross_national_income, data = hdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2884  -2.1655   0.8118   3.1150   9.3923
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                19.5244     2.9684   6.577 6.75e-10 ***
## log_gross_national_income   5.6811     0.3198  17.765  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.721 on 157 degrees of freedom
## Multiple R-squared:  0.6678,    Adjusted R-squared:  0.6657
## F-statistic: 315.6 on 1 and 157 DF,  p-value: < 2.2e-16
```

Exercise 2

An increase in one unit of log gross national income is associated with an increase of 5.6811 years in life expectancy, on average.

A ten percent increase in gross national income is associated with an increase of 5.6811 years in life expectancy, on average.