Basic Regression

Xiaochi

06/01/2020

## One numerical explanatory variable

### Exploratory data analysis

Here are three common steps in an EDA:

1. Most crucially, looking at the raw data values.
2. Computing summary statistics, such as means, medians, and interquartile ranges.
3. Creating data visualizations.

evals\_ch5 <- evals %>%  
 select(ID, score, bty\_avg, age)

evals\_ch5 %>%  
 summarize(mean\_bty\_avg = mean(bty\_avg),   
 mean\_score = mean(score),  
 median\_bty\_avg = median(bty\_avg),   
 median\_score = median(score))

## # A tibble: 1 x 4  
## mean\_bty\_avg mean\_score median\_bty\_avg median\_score  
## <dbl> <dbl> <dbl> <dbl>  
## 1 4.42 4.17 4.33 4.3

evals\_ch5 %>%   
 select(score, bty\_avg) %>%   
 skim()

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 463 |
| Number of columns | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| score | 0 | 1 | 4.17 | 0.54 | 2.30 | 3.80 | 4.30 | 4.6 | 5.00 | ▁▁▅▇▇ |
| bty\_avg | 0 | 1 | 4.42 | 1.53 | 1.67 | 3.17 | 4.33 | 5.5 | 8.17 | ▃▇▇▃▂ |

evals\_ch5 %>%   
 get\_correlation(formula = score ~ bty\_avg)

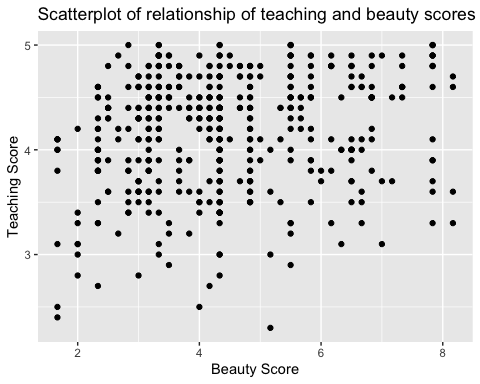
## # A tibble: 1 x 1  
## cor  
## <dbl>  
## 1 0.187

evals\_ch5 %>%   
 summarize(correlation = cor(score, bty\_avg))

## # A tibble: 1 x 1  
## correlation  
## <dbl>  
## 1 0.187

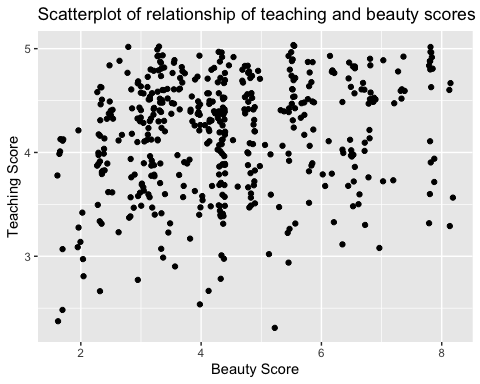
the correlation coefficient of 0.187 indicates that the relationship between teaching evaluation score and “beauty” average is “weakly positive.”

ggplot(evals\_ch5, aes(x = bty\_avg, y = score)) +  
 geom\_point() +  
 labs(x = "Beauty Score",   
 y = "Teaching Score",  
 title = "Scatterplot of relationship of teaching and beauty scores")

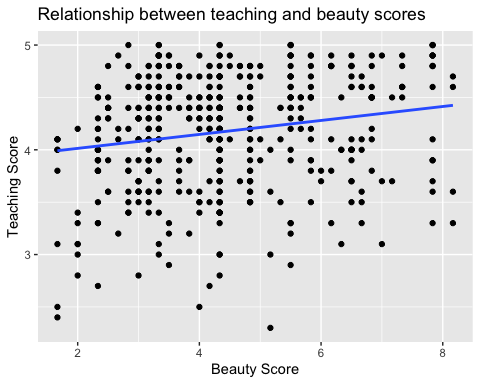


Observe that most “beauty” scores lie between 2 and 8, while most teaching scores lie between 3 and 5. Furthermore, while opinions may vary, it is our opinion that the relationship between teaching score and “beauty” score is “weakly positive.” This is consistent with our earlier computed correlation coefficient of 0.187.

ggplot(evals\_ch5, aes(x = bty\_avg, y = score)) +  
 geom\_jitter() +  
 labs(x = "Beauty Score",   
 y = "Teaching Score",  
 title = "Scatterplot of relationship of teaching and beauty scores")



ggplot(evals\_ch5, aes(x = bty\_avg, y = score)) +  
 geom\_point() +  
 labs(x = "Beauty Score",   
 y = "Teaching Score",  
 title = "Relationship between teaching and beauty scores") +   
 geom\_smooth(method = "lm", se = FALSE)



The positive slope of the blue line is consistent with our earlier observed correlation coefficient of 0.187 suggesting that there is a positive relationship between these two variables: as instructors have higher “beauty” scores, so also do they receive higher teaching evaluations.

While the correlation coefficient and the slope of a regression line always have the same sign (positive or negative), they typically do not have the same value.

evals\_ch5

## # A tibble: 463 x 4  
## ID score bty\_avg age  
## <int> <dbl> <dbl> <int>  
## 1 1 4.7 5 36  
## 2 2 4.1 5 36  
## 3 3 3.9 5 36  
## 4 4 4.8 5 36  
## 5 5 4.6 3 59  
## 6 6 4.3 3 59  
## 7 7 2.8 3 59  
## 8 8 4.1 3.33 51  
## 9 9 3.4 3.33 51  
## 10 10 4.5 3.17 40  
## # … with 453 more rows

evals\_ch5 %>%  
 select(score, age) %>%  
 skim()

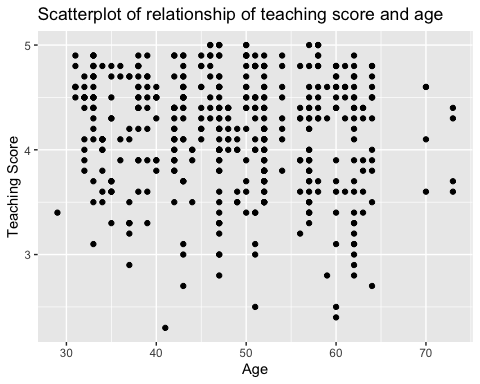
Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 463 |
| Number of columns | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| score | 0 | 1 | 4.17 | 0.54 | 2.3 | 3.8 | 4.3 | 4.6 | 5 | ▁▁▅▇▇ |
| age | 0 | 1 | 48.37 | 9.80 | 29.0 | 42.0 | 48.0 | 57.0 | 73 | ▅▆▇▆▁ |

ggplot(evals\_ch5, aes(x = age, y = score)) +  
 geom\_point() +  
 labs(x = "Age",   
 y = "Teaching Score",  
 title = "Scatterplot of relationship of teaching score and age")



Based on the scatterplot visualization, there seem to have a weak negative relationship between age and teaching score. As age increases, the teaching score see, to decrease slightly.

### Simple linear regression

# Fit regression model:  
score\_model <- lm(score ~ bty\_avg, data = evals\_ch5)  
# Get regression table:  
get\_regression\_table(score\_model)

## # A tibble: 2 x 7  
## term estimate std\_error statistic p\_value lower\_ci upper\_ci  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 intercept 3.88 0.076 51.0 0 3.73 4.03   
## 2 bty\_avg 0.067 0.016 4.09 0 0.035 0.099

The correlation’s interpretation is the “strength of linear association”.

The slope’s interpretation: > For every increase of 1 unit in bty\_avg, there is an associated increase of, on average, 0.067 units of score.

# Fit regression model:  
score\_age\_model <- lm(score ~ age, data = evals\_ch5)  
# Get regression table:  
get\_regression\_table(score\_age\_model)

## # A tibble: 2 x 7  
## term estimate std\_error statistic p\_value lower\_ci upper\_ci  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 intercept 4.46 0.127 35.2 0 4.21 4.71   
## 2 age -0.006 0.003 -2.31 0.021 -0.011 -0.001

For every increase of 1 unit in age, there is an associated decrease of, on average, 0.006 units of score. It matches with the results from our earlier exploratory data analysis.

### Observed/fitted values and residuals

# Fit regression model:  
score\_model <- lm(score ~ bty\_avg, data = evals\_ch5)  
# Get regression table:  
get\_regression\_table(score\_model)

## # A tibble: 2 x 7  
## term estimate std\_error statistic p\_value lower\_ci upper\_ci  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 intercept 3.88 0.076 51.0 0 3.73 4.03   
## 2 bty\_avg 0.067 0.016 4.09 0 0.035 0.099

regression\_points <- get\_regression\_points(score\_model)  
regression\_points

## # A tibble: 463 x 5  
## ID score bty\_avg score\_hat residual  
## <int> <dbl> <dbl> <dbl> <dbl>  
## 1 1 4.7 5 4.21 0.486  
## 2 2 4.1 5 4.21 -0.114  
## 3 3 3.9 5 4.21 -0.314  
## 4 4 4.8 5 4.21 0.586  
## 5 5 4.6 3 4.08 0.52   
## 6 6 4.3 3 4.08 0.22   
## 7 7 2.8 3 4.08 -1.28   
## 8 8 4.1 3.33 4.10 -0.002  
## 9 9 3.4 3.33 4.10 -0.702  
## 10 10 4.5 3.17 4.09 0.409  
## # … with 453 more rows

## One categorical explanatory variable

### Exploratory data analysis

library(gapminder)  
gapminder2007 <- gapminder %>%  
 filter(year == 2007) %>%  
 select(country, lifeExp, continent, gdpPercap)

gapminder2007 %>%  
 select(lifeExp, continent) %>%  
 skim()

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 142 |
| Number of columns | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 1 |
| numeric | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

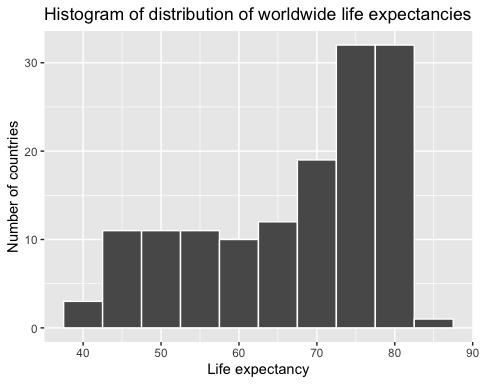
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| continent | 0 | 1 | FALSE | 5 | Afr: 52, Asi: 33, Eur: 30, Ame: 25 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| lifeExp | 0 | 1 | 67.01 | 12.07 | 39.61 | 57.16 | 71.94 | 76.41 | 82.6 | ▂▃▃▆▇ |

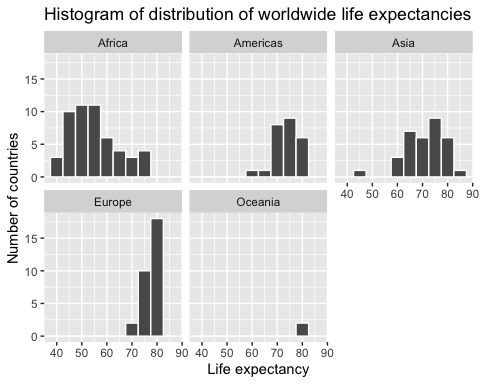
Turning our attention to the summary statistics of the numerical variable lifeExp, we observe that the global median life expectancy in 2007 was 71.94. Thus, half of the world’s countries (71 countries) had a life expectancy less than 71.94. The mean life expectancy of 67.01 is lower, however. Why is the mean life expectancy lower than the median?

ggplot(gapminder2007, aes(x = lifeExp)) +  
 geom\_histogram(binwidth = 5, color = "white") +  
 labs(x = "Life expectancy",   
 y = "Number of countries",  
 title = "Histogram of distribution of worldwide life expectancies")



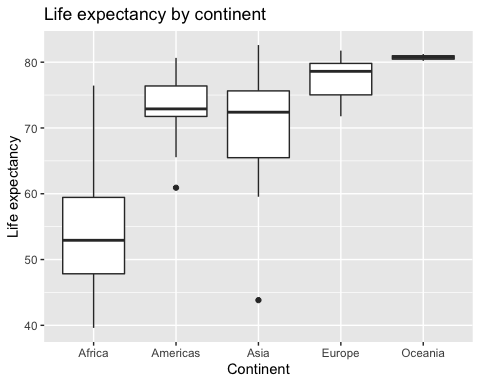
We see that this data is left-skewed, also known as negatively skewed: there are a few countries with low life expectancy that are bringing down the mean life expectancy. However, the median is less sensitive to the effects of such outliers; hence, the median is greater than the mean in this case.

ggplot(gapminder2007, aes(x = lifeExp)) +  
 geom\_histogram(binwidth = 5, color = "white") +  
 labs(x = "Life expectancy",   
 y = "Number of countries",  
 title = "Histogram of distribution of worldwide life expectancies") +  
 facet\_wrap(~ continent, nrow = 2)



Observe that unfortunately the distribution of African life expectancies is much lower than the other continents, while in Europe life expectancies tend to be higher and furthermore do not vary as much. On the other hand, both Asia and Africa have the most variation in life expectancies. There is the least variation in Oceania, but keep in mind that there are only two countries in Oceania: Australia and New Zealand.

ggplot(gapminder2007, aes(x = continent, y = lifeExp)) +  
 geom\_boxplot() +  
 labs(x = "Continent",   
 y = "Life expectancy",  
 title = "Life expectancy by continent")



lifeExp\_by\_continent <- gapminder2007 %>%  
 group\_by(continent) %>%  
 summarize(median = median(lifeExp),   
 mean = mean(lifeExp))

Observe the order of the second column median life expectancy: Africa is lowest, the Americas and Asia are next with similar medians, then Europe, then Oceania. This ordering corresponds to the ordering of the solid black lines inside the boxes in our side-by-side boxplot in Figure 5.9.

gapminder2007

## # A tibble: 142 x 4  
## country lifeExp continent gdpPercap  
## <fct> <dbl> <fct> <dbl>  
## 1 Afghanistan 43.8 Asia 975.  
## 2 Albania 76.4 Europe 5937.  
## 3 Algeria 72.3 Africa 6223.  
## 4 Angola 42.7 Africa 4797.  
## 5 Argentina 75.3 Americas 12779.  
## 6 Australia 81.2 Oceania 34435.  
## 7 Austria 79.8 Europe 36126.  
## 8 Bahrain 75.6 Asia 29796.  
## 9 Bangladesh 64.1 Asia 1391.  
## 10 Belgium 79.4 Europe 33693.  
## # … with 132 more rows

gapminder2007 %>%  
 select(gdpPercap, continent) %>%  
 skim()

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 142 |
| Number of columns | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 1 |
| numeric | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

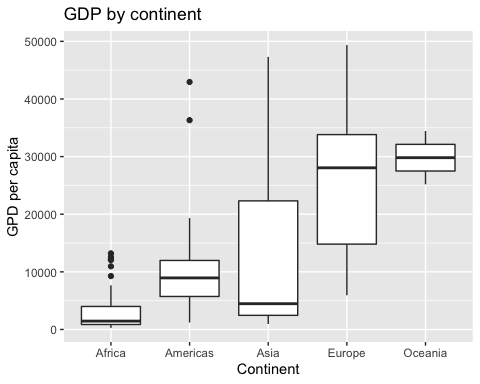
**Variable type: factor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| continent | 0 | 1 | FALSE | 5 | Afr: 52, Asi: 33, Eur: 30, Ame: 25 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| gdpPercap | 0 | 1 | 11680.07 | 12859.94 | 277.55 | 1624.84 | 6124.37 | 18008.84 | 49357.19 | ▇▂▁▂▁ |

ggplot(gapminder2007, aes(x = continent, y = gdpPercap)) +  
 geom\_boxplot() +  
 labs(x = "Continent",   
 y = "GPD per capita",  
 title = "GDP by continent")



Based on this exploration, it seems that GDP’s are very different among different continents, which means that continent might be a statistically significant predictor for an area’s GDP.

### Linear regression

lifeExp\_by\_continent <- gapminder2007 %>%  
 group\_by(continent) %>%  
 summarize(median = median(lifeExp),   
 mean = mean(lifeExp))

lifeExp\_model <- lm(lifeExp ~ continent, data = gapminder2007)  
get\_regression\_table(lifeExp\_model)

## # A tibble: 5 x 7  
## term estimate std\_error statistic p\_value lower\_ci upper\_ci  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 intercept 54.8 1.02 53.4 0 52.8 56.8  
## 2 continentAmericas 18.8 1.8 10.4 0 15.2 22.4  
## 3 continentAsia 15.9 1.65 9.68 0 12.7 19.2  
## 4 continentEurope 22.8 1.70 13.5 0 19.5 26.2  
## 5 continentOceania 25.9 5.33 4.86 0 15.4 36.4

To summarize, the 5 values in the estimate column in Table 5.8 correspond to the “baseline for comparison” continent Africa (the intercept) as well as four “offsets” from this baseline for the remaining 4 continents: the Americas, Asia, Europe, and Oceania.

# Fit regression model:  
gdp\_model <- lm(gdpPercap ~ continent, data = gapminder2007)  
# Get regression table:  
get\_regression\_table(gdp\_model)

## # A tibble: 5 x 7  
## term estimate std\_error statistic p\_value lower\_ci upper\_ci  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 intercept 3089. 1373. 2.25 0.026 375. 5804.  
## 2 continentAmericas 7914. 2409. 3.28 0.001 3150. 12678.  
## 3 continentAsia 9384. 2203. 4.26 0 5027. 13741.  
## 4 continentEurope 21965. 2270. 9.68 0 17478. 26453.  
## 5 continentOceania 26721. 7133. 3.75 0 12616. 40826.

In our previous exploratory data analysis, it seemed that continent is a statistically significant predictor for an area’s GDP. Here, by fit a new linear regression using lm(gdpPercap ~ continent, data = gapminder2007) where gdpPercap is the new outcome variable  
y, we are able to write an equation to predict gdpPercap using the continent as statistically significant predictors. Therefore, the regression results matches with the results from your previous exploratory data analysis.

### Observed/fitted values and residuals

lifeExp\_model <- lm(lifeExp ~ continent, data = gapminder2007)  
get\_regression\_table(lifeExp\_model)

## # A tibble: 5 x 7  
## term estimate std\_error statistic p\_value lower\_ci upper\_ci  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 intercept 54.8 1.02 53.4 0 52.8 56.8  
## 2 continentAmericas 18.8 1.8 10.4 0 15.2 22.4  
## 3 continentAsia 15.9 1.65 9.68 0 12.7 19.2  
## 4 continentEurope 22.8 1.70 13.5 0 19.5 26.2  
## 5 continentOceania 25.9 5.33 4.86 0 15.4 36.4

regression\_points <- get\_regression\_points(lifeExp\_model, ID = "country")  
regression\_points

## # A tibble: 142 x 5  
## country lifeExp continent lifeExp\_hat residual  
## <fct> <dbl> <fct> <dbl> <dbl>  
## 1 Afghanistan 43.8 Asia 70.7 -26.9   
## 2 Albania 76.4 Europe 77.6 -1.23   
## 3 Algeria 72.3 Africa 54.8 17.5   
## 4 Angola 42.7 Africa 54.8 -12.1   
## 5 Argentina 75.3 Americas 73.6 1.71   
## 6 Australia 81.2 Oceania 80.7 0.515  
## 7 Austria 79.8 Europe 77.6 2.18   
## 8 Bahrain 75.6 Asia 70.7 4.91   
## 9 Bangladesh 64.1 Asia 70.7 -6.67   
## 10 Belgium 79.4 Europe 77.6 1.79   
## # … with 132 more rows

Using the sorting functionality of RStudio’s spreadsheet viewer, we can identify that the five countries with the five smallest (most negative) residuals are: Afghanistan, Swaziland, Mozambique, Haiti, and Zambia. These negative residuals indicate that these data points have the biggest negative deviations from their group means. This means that these five countries’ average life expectancies are the lowest comparing to their respective continents’ average life expectancies. For example, the residual for Afghanistan is −26.900 and it is the smallest residual. This means that the average life expectancy of Afghanistan is  
26.900 years lower than the average life expectancy of its continent, Asia.

Using either the sorting functionality of RStudio’s spreadsheet viewer, we can identify that the five countries with the five largest (most positive) residuals are: Reunion, Libya, Tunisia, Mauritius, and Algeria. These positive residuals indicate that the data points are above the regression line with the longest distance. This means that these five countries’ average life expectancies are the highest comparing to their respective continents’ average life expectancies. For example, the residual for Reunion is 21.636 and it is the largest residual. This means that the average life expectancy of Reunion is  
21.636 years higher than the average life expectancy of its continent, Africa.

## Related topics

### Correlation is not necessarily causation

### Best-fitting line

# Fit regression model:  
score\_model <- lm(score ~ bty\_avg, data = evals\_ch5)  
  
# Get regression points:  
regression\_points <- get\_regression\_points(score\_model)  
regression\_points

## # A tibble: 463 x 5  
## ID score bty\_avg score\_hat residual  
## <int> <dbl> <dbl> <dbl> <dbl>  
## 1 1 4.7 5 4.21 0.486  
## 2 2 4.1 5 4.21 -0.114  
## 3 3 3.9 5 4.21 -0.314  
## 4 4 4.8 5 4.21 0.586  
## 5 5 4.6 3 4.08 0.52   
## 6 6 4.3 3 4.08 0.22   
## 7 7 2.8 3 4.08 -1.28   
## 8 8 4.1 3.33 4.10 -0.002  
## 9 9 3.4 3.33 4.10 -0.702  
## 10 10 4.5 3.17 4.09 0.409  
## # … with 453 more rows

# Compute sum of squared residuals  
regression\_points %>%  
 mutate(squared\_residuals = residual^2) %>%  
 summarize(sum\_of\_squared\_residuals = sum(squared\_residuals))

## # A tibble: 1 x 1  
## sum\_of\_squared\_residuals  
## <dbl>  
## 1 132.

### get\_regression\_x() functions

# Fit regression model:  
score\_model <- lm(formula = score ~ bty\_avg, data = evals\_ch5)  
# Get regression table:  
get\_regression\_table(score\_model)

## # A tibble: 2 x 7  
## term estimate std\_error statistic p\_value lower\_ci upper\_ci  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 intercept 3.88 0.076 51.0 0 3.73 4.03   
## 2 bty\_avg 0.067 0.016 4.09 0 0.035 0.099

library(broom)  
library(janitor)

##   
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

score\_model %>%  
 tidy(conf.int = TRUE) %>%  
 mutate\_if(is.numeric, round, digits = 3) %>%  
 clean\_names() %>%  
 rename(lower\_ci = conf\_low, upper\_ci = conf\_high)

## # A tibble: 2 x 7  
## term estimate std\_error statistic p\_value lower\_ci upper\_ci  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 3.88 0.076 51.0 0 3.73 4.03   
## 2 bty\_avg 0.067 0.016 4.09 0 0.035 0.099

score\_model %>%  
 augment() %>%  
 mutate\_if(is.numeric, round, digits = 3) %>%  
 clean\_names() %>%  
 select(-c("se\_fit", "hat", "sigma", "cooksd", "std\_resid"))

## # A tibble: 463 x 4  
## score bty\_avg fitted resid  
## <dbl> <dbl> <dbl> <dbl>  
## 1 4.7 5 4.21 0.486  
## 2 4.1 5 4.21 -0.114  
## 3 3.9 5 4.21 -0.314  
## 4 4.8 5 4.21 0.586  
## 5 4.6 3 4.08 0.52   
## 6 4.3 3 4.08 0.22   
## 7 2.8 3 4.08 -1.28   
## 8 4.1 3.33 4.10 -0.002  
## 9 3.4 3.33 4.10 -0.702  
## 10 4.5 3.17 4.09 0.409  
## # … with 453 more rows