Math 300 Lesson 12 Notes

Simple Linear Regression - Discrete Predictor

YOUR NAME HERE

June, 2022

Contents

Objectives

1. Explore the relationship between 2 variables, one numerical and one categorical, using summary statistics and visualizations in `R`.
2. Fit a linear regression model to two variables, one numerical response and one categorical predictor, using the `lm()` function and interpret the output. This includes the interpretation of baseline mean and offsets.
3. Generate a table of observations, fitted values, and residuals from a linear regression object.

Reading

Chapter 5.2

Lesson

Work through the learning checks LC5.4 - LC5.7. Complete the code as necessary.

- The response `y` is the numeric variable. Math 378 discusses cases where the response is categorical. Understanding the regression output here is important. There is no line just a baseline average and offsets from that.
- The regression output will still predict the **mean** value of the response variable.
- The baseline is an **average** and is the first level of the factor based on alphabetic order.

Setup

```
library(tidyverse)
library(moderndive)
```

```
## Warning: package 'moderndive' was built under R version 4.1.3
```

```
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 4.1.3
```

```
library(gapminder)
```

```
## Warning: package 'gapminder' was built under R version 4.1.3
```

Create the data needed for the exercises.

```
# Complete the code and remove comment labels
#gapminder2007 <- _________ %>%
#  filter(year == _______) %>%
#  select(country, lifeExp, ________, gdpPercap)
```

Let's look at 5 random rows of data.

```
set.seed(1234)
gapminder2007 %>%
  sample_n(size = 5)
```

LC 5.4 (Objective 1)

(LC 5.4) Conduct a new exploratory data analysis with the same explanatory variable x being `continent` but with `gdpPercap` as the new outcome variable y. Remember, this involves three things:

- Most crucially: Looking at the raw data values.
- Computing summary statistics, such as means, medians, and interquartile ranges.
- Creating data visualizations.

What can you say about the differences in GDP per capita between continents based on this exploration?

Solution:

- Looking at the raw data values:

```
# Complete the code and remove comment labels
#_______________(gapminder2007)
```

- Computing summary statistics, such as means, medians, and interquartile ranges:

```
gapminder2007 %>%
  select(gdpPercap, continent) %>%
  my_skim()
```

- Creating data visualizations:

Create boxplots

```
# Complete the code and remove comment labels
#ggplot(gapminder2007, aes(x = ______________, y = gdpPercap)) +
#  geom_XXXXXX() +
#  labs(
#    x = "Continent", y = "GPD per capita",
#    title = "_________________") +
#  theme_bw()
```

LC 5.5 (Objective 2)

(LC 5.5) Fit a new linear regression using `lm(gdpPercap ~ continent, data = gapminder2007)` where `gdpPercap` is the new outcome variable y. Get information about the "best-fitting" line from the regression table by applying the `get_regression_table()` function. How do the regression results match up with the results from your previous exploratory data analysis?

Solution:

LC 5.6 (Objective 3)

(LC 5.6) Using either the sorting functionality of RStudio's spreadsheet viewer or using the data wrangling tools you learned in Chapter @ref(wrangling), identify the five countries with the five smallest (most negative) residuals? What do these negative residuals say about their life expectancy relative to their continents?

Solution:

We switched by to life expectancy. We need the model.

```
# Complete the code and remove comment labels
# lifeExp_model <- lm(_________ ~ continent, data = gapminder2007)
```

Use `R`.

```
# Complete the code and remove comment labels
#get_regression_points(lifeExp_model, ID = "_____________") %>%
#  arrange(__________) %>%
#  slice_head(n=_______)
```

LC 5.7 (Objective 3)

(LC 5.7) Repeat this process, but identify the five countries with the five largest (most positive) residuals. What do these positive residuals say about their life expectancy relative to their continents?

Solution:

Using `R`.

Documenting software

- File creation date: 2022-06-23
- R version 4.1.1 (2021-08-10)
- `tidyverse` package version: 1.3.1
- `skimr` package version: 2.1.4
- `gapminder` package version: 0.3.0

- `moderndive` package version: 0.5.4