

Math 300 NTI Lesson 12

Simple Linear Regression - Discrete x

Professor Bradley Warner

June, 2022

Contents

Objectives	1
Reading	1
Lesson	1
Documenting software	6

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 and one categorical, 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

Remember that you will be running this more like a lab than a lecture. You want them using R and answering questions. Have them open the notes rmd and work through it together.

Work through the learning checks LC5.4 - LC5.7.

- 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(moderndiver)
library(skimr)
library(gapminder)
```

Create the data needed for the exercises.

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

Let's look at 5 random rows of data.

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

```
## # A tibble: 5 x 4
##   country      lifeExp continent gdpPercap
##   <fct>        <dbl> <fct>      <dbl>
## 1 Congo, Dem. Rep.  46.5 Africa      278.
## 2 Mali           54.5 Africa     1043.
## 3 Peru           71.4 Americas    7409.
## 4 Senegal        63.1 Africa     1712.
## 5 Venezuela      73.7 Americas   11416.
```

LC 5.4 (Objective 1)

(LC5.4) 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:

```
glimpse(gapminder2007)
```

```
## Rows: 142
## Columns: 4
## $ country   <fct> "Afghanistan", "Albania", "Algeria", "Angola", "Argentina", ~
## $ lifeExp   <dbl> 43.828, 76.423, 72.301, 42.731, 75.320, 81.235, 79.829, 75.6~
## $ continent <fct> Asia, Europe, Africa, Africa, Americas, Oceania, Europe, Asi~
## $ gdpPercap <dbl> 974.5803, 5937.0295, 6223.3675, 4797.2313, 12779.3796, 34435~
```

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

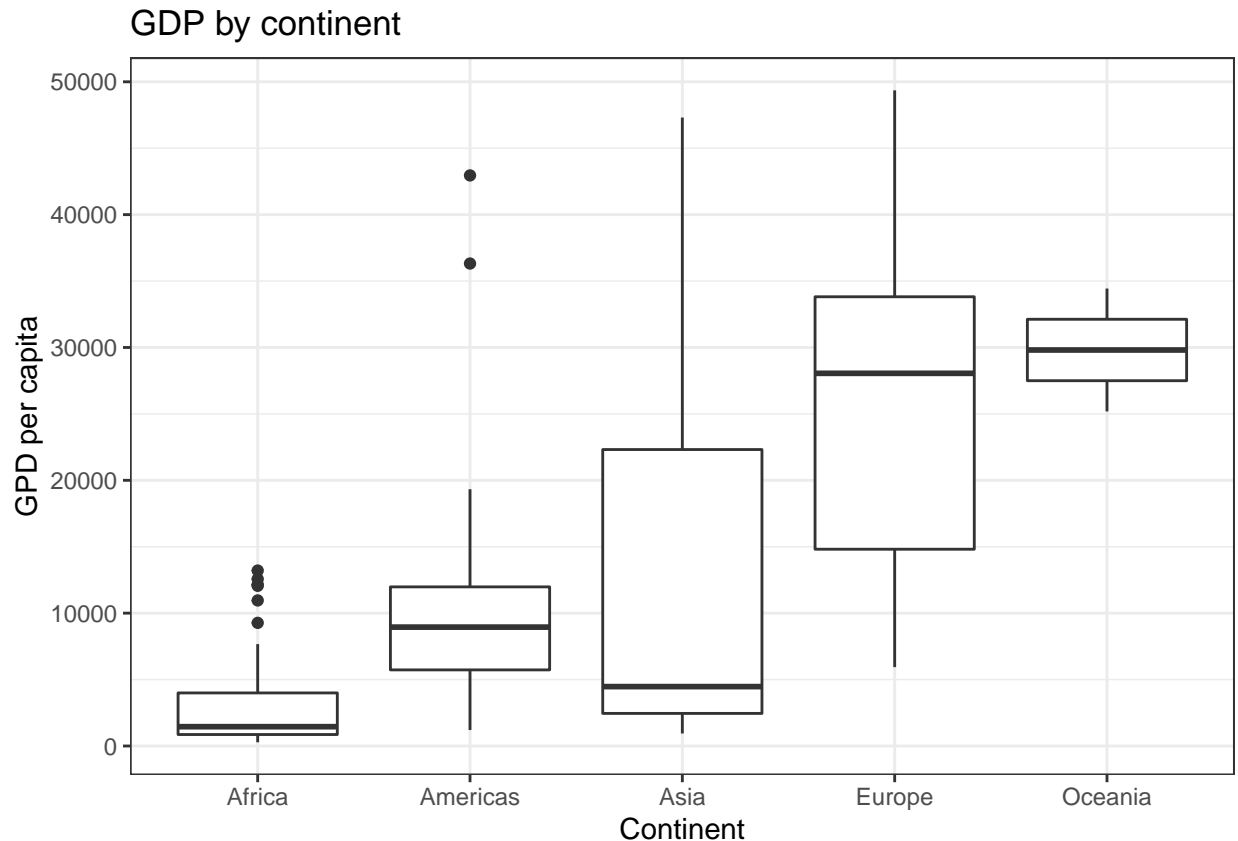
```
gapminder2007 %>%
  select(gdpPercap, continent) %>%
  my_skim() %>%
  print()

## -- Data Summary -----
##                               Values
## 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
## 1 continent      0              1 FALSE           5
##   top_counts
## 1 Afr: 52, Asi: 33, Eur: 30, Ame: 25
##
## -- Variable type: numeric -----
##   skim_variable n_missing complete_rate  mean    sd  p0  p25  p50  p75
## 1 gdpPercap      0              1 11680. 12860. 278. 1625. 6124. 18009.
##   p100
## 1 49357.

## $factor
##
## -- Variable type: factor -----
##   skim_variable n_missing complete_rate ordered n_unique top_counts
## 1 continent      0              1 FALSE           5 Afr: 52, Asi: 33, Eur:~
##
## $numeric
##
## -- Variable type: numeric -----
##   skim_variable n_missing complete_rate  mean    sd  p0  p25  p50  p75
## 1 gdpPercap      0              1 11680. 12860. 278. 1625. 6124. 18009.
## # ... with 1 more variable: p100 <dbl>
```

- Creating data visualizations:

```
ggplot(gapminder2007, aes(x = continent, y = gdpPercap)) +
  geom_boxplot() +
  labs(
    x = "Continent", y = "GPD per capita",
    title = "GDP by continent") +
  theme_bw()
```



Based on this exploration, it seems that GDP's are very different among different continents, in terms of medians, variation, and symmetry. At a minimum this means that continent might be an important predictor for an area's **mean** GDP.

LC 5.5 (Objective 2)

(LC5.5) 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:

```
# 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 continent: Americas  7914.    2409.     3.28  0.001   3150.  12678.
## 3 continent: Asia     9384.    2203.     4.26   0       5027.  13741.
```

```
## 4 continent: Europe      21965.      2270.      9.68  0      17478.      26453.
## 5 continent: Oceania    26721.      7133.      3.75  0      12616.      40826.
```

$$\begin{aligned}\widehat{y} = \widehat{\text{gdpPercap}} &= b_0 + b_{\text{Amer}} \cdot \mathbb{K}_{\text{Amer}}(x) + b_{\text{Asia}} \cdot \mathbb{K}_{\text{Asia}}(x) + \\ &\quad b_{\text{Euro}} \cdot \mathbb{K}_{\text{Euro}}(x) + b_{\text{Ocean}} \cdot \mathbb{K}_{\text{Ocean}}(x) \\ &= 3089 + 7914 \cdot \mathbb{K}_{\text{Amer}}(x) + 9384 \cdot \mathbb{K}_{\text{Asia}}(x) + \\ &\quad 21965 \cdot \mathbb{K}_{\text{Euro}}(x) + 26721 \cdot \mathbb{K}_{\text{Ocean}}(x)\end{aligned}$$

In our previous exploratory data analysis, it seemed that continent is a statistically significant predictor for an area's **mean** GDP. Here, by fitting 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 **average gdpPercap** using the continent as a predictor. Therefore, the regression results matches with the results from your previous exploratory data analysis.

LC 5.6 (Objective 3)

(LC5.6) 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.

```
lifeExp_model <- lm(lifeExp ~ continent, data = gapminder2007)
```

Let's use R.

```
get_regression_points(lifeExp_model, ID = "country") %>%
  arrange(residual) %>%
  slice_head(n=5)
```

```
## # A tibble: 5 x 5
##   country      lifeExp continent lifeExp_hat residual
##   <fct>        <dbl> <fct>         <dbl>     <dbl>
## 1 Afghanistan  43.8 Asia          70.7     -26.9
## 2 Swaziland    39.6 Africa         54.8     -15.2
## 3 Mozambique  42.1 Africa         54.8     -12.7
## 4 Haiti       60.9 Americas       73.6     -12.7
## 5 Zambia      42.4 Africa         54.8     -12.4
```

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 compared to their respective continents' average life expectancies. For example, the residual for Afghanistan is -26.9 and it is the smallest residual. This means that the average life expectancy of Afghanistan is 26.9 years lower than the average life expectancy of its continent, Asia.

LC 5.7 (Objective 3)

(LC5.7) 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.

```
get_regression_points(lifeExp_model, ID = "country") %>%  
  arrange(desc(residual)) %>%  
  slice_head(n=5)
```

```
## # A tibble: 5 x 5  
##   country lifeExp continent lifeExp_hat residual  
##   <fct>      <dbl> <fct>          <dbl>    <dbl>  
## 1 Reunion    76.4 Africa         54.8     21.6  
## 2 Libya      74.0 Africa         54.8     19.1  
## 3 Tunisia    73.9 Africa         54.8     19.1  
## 4 Mauritius  72.8 Africa         54.8     18.0  
## 5 Algeria    72.3 Africa         54.8     17.5
```

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.

Documenting software

- File creation date: 2022-06-04
- R version 4.1.3 (2022-03-10)
- tidyverse package version: 1.3.1
- skimr package version: 2.1.4
- gapminder package version: 0.3.0
- moderndive package version: 0.5.4