

R Practice 2: ggplot2 and Simple Regression

Ryan Safner

October 3, 2018

1. Install and load the package `gapminder`. Type `?gapminder` and hit enter to see a description of the data.
2. Get summary statistics of `gdpPercap`.
3. Use base R's `hist()` function to plot a histogram of `gdpPercap`
4. Now load and use `ggplot2` to create a histogram of `gdpPercap`. Remember your base layer must establish which data frame you are using (`gapminder`) and the base aesthetics `aes()` to define what variable is `x`. Your second layer is a `geom_histogram()`
5. Get summary statistics of `lifeExp`.
6. Use base R's `hist()` function to create a histogram of `lifeExp`.
7. Use `ggplot2` to create a histogram of `lifeExp`.
8. Instead of a histogram, make a density plot of `lifeExp` with `geom_density()`
9. Using base R's `boxplot()` function, create a boxplot of `gdpPercap` by continent.
10. Now do the same with `ggplot2`. In your initial aesthetics, set `x` as `continent`, `y` as `gdpPercap` and `fill (color)` by `continent`. Your `geom` layer is `geom_boxplot()`.
11. The nice thing about building plots one layer at a time is that we can use different `geoms` on the same base layer. Replicate your answer to #10 and instead of `geom_boxplot()`, try a "Violin plot" with `geom_violin()`.
12. Use what you've learned so far to make a density plot of `gdpPercap` by continent. Note your only variable here is `x`. Add an option to your `geom_density` layer of setting `alpha=0.5` (to make plots more transparent).
13. Do the same thing for `lifeExp`
14. Now let's try to estimate the following relationship.

$$\widehat{\text{Life Expectancy}} = \beta_0 + \beta_1 \text{GDP Per Capita}$$

First, use base R to make a scatterplot of these two variables with `plot()`. Be sure to signify `x` and `y` using the `data.frame$variable` syntax.

15. Now let's try with `ggplot2`. For your base layer, consider in your aesthetics what is `x` and what is `y`. We want our data to manifest as data points, so use `geom_point()` as your second layer. Be sure to save this as some object.
16. Now on top of the existing plot, let's add a regression line. Redefine your object to be itself `+geom_smooth(method="lm")` to add the regression line (`geom_smooth` creates a smooth line, and `lm` stands for linear model, i.e. OLS regression).
17. Now let's spice this up a bit. Recreate your plot but this time, include in your base layer's aesthetics (in addition to defining `x` and `y`) `color=continent` to color by continent.
18. Now add a regression line. Notice that since we initially defined in the base layer to color by continent, it also creates different colored lines, one for each continent.
19. Let's try facetting. Add to your previous plot `+facet_grid(cols=vars(continent))`. This creates a grid of individual plots, one for each continent, and arranges them into columns (`cols`) by the variable `continent`.
20. Let's try only looking at the year 2002. We can use the `subset()` function to create another data frame for only the year 2000 like `gapminder.2002<-subset(gapminder, year==2002)`. Next, get summary statistics for the gdp per capita in 2002.
21. Plot a histogram of gdp per capita in `ggplot2` for 2002
22. Plot a scatterplot with `ggplot2` for 2002 gdp per capita (`x`) vs. life expectancy (`y`)
23. Now let's add more information to our scatterplot. Add an option to the `geom_point()` to plot `size=pop`.

Regression Analysis

24. Now let's turn away from data visualization to more technical analysis with regression. For more information and examples, see lecture 7. Run a regression of life expectancy on gdp per capita. `summary()` your regression. What are:
 - $\hat{\beta}_0$
 - $\hat{\beta}_1$
 - $SE(\hat{\beta}_0)$
 - $SE(\hat{\beta}_1)$
 - R^2
 - SER
25. Is $\hat{\beta}_1$ statistically significantly different from 0 (i.e. $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$)? How do you know? See lecture 8 for more help.
26. Save the residuals and plot them in a residual plot (using the residuals as `y` instead of `lifeExp`). Add a horizontal line at 0 with `geom_vline(yintercept=0)`
27. Install and then load `stargazer` to output your regression into a table. For simplicity, set `type=text` for now. Verify where everything is that you found for question #24.