Assignment 4

Due: December 8, 2017 @ 11:59pm

Instructions

In the Github Classroom repository that you copied, you should have a template file named `assignment_4.Rmd`. Use this file to answer the questions below. Also fill out any missing information at the top of the RMarkdown document. Once you're done, be sure to save, commit, and push to Github. Then, submit a pull request.

Questions

Starbucks and linear modeling

The CSV file `starbucks.csv` is a dataset containing the nutritional information of the food items sold at Starbucks.[1] Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbs a menu item has based on its calorie content.

1. Before creating the linear model, based on your basic knowledge of nutrition, what is the relationship between calories and the amount of carbs a food item has? For these proposed variables, which is the *explanatory variable* and which is the *response variable*? Do you expect that such a model would be valid for any amount of calories (within statistical error)? Why or why not?

2. Using `lm()`, create a linear model of the relationship between the number of calories and amount of carbohydrates (in grams) in the Starbucks food menu items. Plot the linear model over top of the data points. Visualize the residuals using a scatter plot and a *frequency polygon* plot.

3. Do these data meet the conditions required for fitting a least-squares line, and can we call this a good model? If it doesn't meet the conditions for a least-squares line, what important information might we be missing in our model?

   Hint: Look at the other information that's available in the dataset.

Least-squares linear modeling versus other choices

The `lm()` function in R takes a dataset and creates a linear model using least-squares regression. This is often a useful choice, but one downside is that these models are sensitive to unusual values because squaring distances place more of an emphasis on larger residuals. It isn't a strict requirement that we use this criteria, so let's analyze what outliers can do to a least-squares linear model and then work with a couple of alternative fitting methods.

4. The dataset in `fitting_datasets.rds` consists of three columns, `x` and `y`, which are the main dataset, and `label`, which splits the data into three categories labelled `"set_a"`, `"set_b"`, and `"set_c"`. Filter the data so that you can fit a linear model to the subsetted data **for each of the three label categories**. Fit the linear model and visualize the results. Then, verify whether the a linear model is appropriate in each case by checking that the three conditions are met: *linearity*, *nearly normal residuals*, and *constant variability* This will require examining the residuals[2] and making visualizations. State which, if any, of these datasets should not be modeled as a linear function.

5. An alternative to the least-squares criterion is the mean-absolute distance criterion, which involves averaging over the *absolute value* of residuals instead of squaring them. This can be implemented by using the function `optim()` in combination with the following custom function (type the custom function below into your RMarkdown file exactly):

[1] Dataset source: Starbucks.com, collected on March 10, 2011, http://www.starbucks.com/menu/nutrition.

[2] For assistance, consult the procedure for creating the figure at the end of Chapter 23.3 in *R for Data Science*.

```
make_prediction <- function(intercept_slope_parameters, data) {
  intercept_slope_parameters[1] + data$x * intercept_slope_parameters[2]
}

measure_distance <- function(intercept_slope_parameters, dataset) {
  diff <- dataset$y - make_prediction(intercept_slope_parameters, dataset)

  mean(abs(diff))
}
```

After typing in the above functions, remind yourself how `optim()` is used by reviewing Chapters 23.2 and 23.3 of *R for Data Science*. Then, use `optim()` to redo the fits of `"set_a"`, `"set_b"`, and `"set_c"` using the mean-absolute distance and visualize the results, including the residuals. Then, **for** `"set_a"` **only**, create a plot that overlays the mean-absolute distance fit (this question) and least-squares fit (previous question) lines on a scatterplot of the `"set_a"` data points. Also create a histogram plot that overlays the `"set_a"` residuals from both types of linear models for comparison. Compare the two types of linear models and note any differences that you observe between the two.