AE 14: Model Comparison

Tips Data

Driver: _____, Reporter: _____, Gopher: _____

> **Important**
>
> - Open RStudio and create a subfolder in your AE folder called "AE-14".
> - Go to the Canvas and locate your `AE-14` assignment to get started.
> - Upload the `ae-14.qmd` and `tip-data.csv` files into the folder you just created. The `.qmd` and PDF responses are due in Canvas. You can check the due date on the Canvas assignment.

Packages + data

```
library(tidyverse)
library(broom)
library(yardstick)
library(ggformula)
library(patchwork)
library(knitr)
library(kableExtra)

tips <- read_csv("tip-data.csv")
```

What factors are associated with the amount customers tip at a restaurant? To answer this question, we will use data collected in 2011 by a student at St. Olaf who worked at a local restaurant.[1]

The variables we'll focus on for this analysis are

[1] Dahlquist, Samantha, and Jin Dong. 2011. "The Effects of Credit Cards on Tipping." Project for Statistics 212-Statistics for the Sciences, St. Olaf College.

- `Tip`: amount of the tip
- `Meal`: which meal this was (`Lunch`, `Dinner`, `Late Night`)
- `Party`: number of people in the party
- `Age`: Age category of person paying the bill (`Yadult`, `Middle`, `SenCit`)

Analysis goal

The goals of this activity are to:

- Use ANOVA to determine whether our model is useful as a whole
- Begin thinking about R^2 in a multivariate setting

Exercise 0

Complete the following to clean and then observe your data:

1. Use `drop_na` to remove any rows where `Party` is missing.

```
tips <- tips |>
  _______   # drop missing values from party
```

2. Generate a bar chart of the variable `Meal`.

```
# insert code here
```

3. Run the following code and generate the same bar chart as above. You can even copy and paste your code. What's the difference between the two plots? What do you think `fct_relevel` does?

```
tips <- tips |>
  mutate(
    Meal = fct_relevel(Meal, "Lunch", "Dinner", "Late Night"),
    Age  = fct_relevel(Age, "Yadult", "Middle", "SenCit")
  )

# Generate plot here
```

Exercise 1

Fit a linear model to predict `Tips` from `Party` and `Age`.

Exercise 2

Pipe the model you generated in the previous Exercise into the function `anova`.

Exercise 3

Based on the output above, compute $SSTotal$, $SSError$, and $SSModel$. You should only need to use addition and/or subtraction.

Exercise 4

Pipe your model into the `glance` function. Identify, the F-statistic and p-value for an F-test of this model. Interpret the outcome of your test in the context of the problem. Be prepared to discuss the difference between this p-value and the p-values from the individual model coefficients.

Exercise 5

What is the R^2 value for this model?

Exercise 6

Fit the full model. What is it's R^2? Does this model have a higher or lower R^2 than the previous model? Does this mean it is a better model? Be prepared to discuss.

Exercise 7

The following code converts the numerical variable `Bill` into a new categorical variable `Bill_factor`. Essentially, each different number in `Bill` is treated as it's own category. Fit a model predicting `Tip` from `Bill_factor`. What is your `R^2`? Think about the implications of this and what it means for the usefullness of R^2.

```
tips <- tips |>
  mutate(Bill_factor = factor(Bill))
```

To submit the AE

> **!** Important
>
> - Render the document to produce the PDF with all of your work from today's class.
> - Upload your QMD and PDF files to the Canvas assignment.