

# AE 4: Exam 1 Review

Add your name here

## Packages

```
library(tidyverse)
library(tidymodels)
library(ggfortify)
library(knitr)
```

## Restaurant tips

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.<sup>1</sup>

The variables we'll focus on for this analysis are

- Tip: amount of the tip
- Party: number of people in the party

View the data set to see the remaining variables.

```
tips <- read_csv("data/tip-data.csv") %>%
  mutate(obs_num = row_number())
```

## Exploratory analysis

1. Visualize, summarize, and describe the relationship between **Party** and **Tips**.

---

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

```
# add code
```

## Modeling

Let's start by fitting a model using **Party** to predict the **Tips** at this restaurant.

2. Write the statistical model.
3. Fit the regression line and write the regression equation. Name the model `tips_fit` and display the results with `kable()` and a reasonable number of digits.

```
# add code
```

4. Interpret the slope.
5. Does it make sense to interpret the intercept? Explain your reasoning.

## Inference

### Inference for the slope

6. Construct a 90% confidence interval for the slope using bootstrapping and the percentile method and interpret it in context of the data.

```
# add code
```

7. Conduct a hypothesis test at the equivalent significance level using permutation. State the hypotheses and the significance level you're using explicitly.

```
# add code
```

8. Check the relevant conditions for Exercises 7 and 8. Are there any violations in conditions that make you reconsider your inferential findings?

```
# add code
```

9. Now repeat Exercises 7 and 8 using approaches based on mathematical models.

```
# add code
```

10. Check the relevant conditions for Exercise 9. Are there any violations in conditions that make you reconsider your inferential findings?

## Inference for a prediction

11. Predict the tip for a party of 4.

```
# add code
```

12. Suppose you're asked to construct a confidence and a prediction interval for your finding in Exercise 11. Which one would you expect to be wider and why? In your answer clearly state the difference between these intervals.
13. Now construct the intervals from Exercise 12 and comment on whether your guess is confirmed.

```
# add code
```

## Model diagnostics

### Leverage (Outliers in x direction)

14. What is the threshold used to identify observations with high leverage? Calculate the threshold and save the value as `leverage_threshold`.

```
## add code
```

15. Make a plot of the standardized residuals vs. leverage (`which = 5`). Use `geom_vline()` to add a vertical line to help identify points with high leverage.

```
## add code
```

16. Let's dig into the data further. Which observations have high leverage? Why do these points have high leverage?

```
## add code
```

### Identifying outliers (outliers in y direction)

17. Make a plot of the residuals vs. fitted values a plot of the `sqrt(standardized residuals)` vs. fitted (You can use `which = c(1, 3)` to display the plots side-by-side).

- How are the plots similar? How do they differ?
- What is an advantage of using the plot of the residuals vs. fitted to check conditions and model diagnostics?
- What is an advantage of using the plot of the `sqrt(standardized residuals)` vs. fitted to check conditions and model diagnostics?

```
# add code here
```

18. Are there any observations that are outliers?

### Cook's distance

19. Make a plot to check Cook's distance (`which = 4`). Based on this plot, are there any points that have a strong influence on the model coefficients?

```
## add code
```

### Adding another variable

20. Add another variable, `Payment`, to your exploratory visualization. Describe any patterns that emerge.

```
# add code
```

21. Fit a multiple linear regression model predicting `Tip` from `Party` and `Payment`. Display the results with `kable()` and a reasonable number of digits.

```
# add code
```

21. Interpret the slopes.
22. Does it make sense to interpret the intercept? Explain your reasoning.