AE 4: Exam 1 Review

Add your name here

Packages

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

knitr::opts_chunk$set(
  fig.asp = 0.618,
  out.width = "80%"
)
```

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.[1]

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.

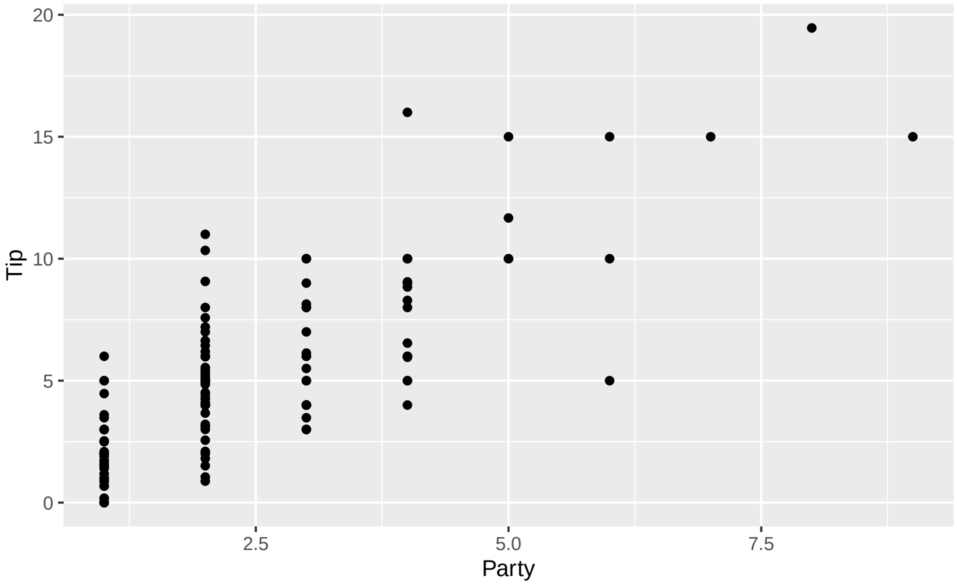
```r
tips <- read_csv("data/tip-data.csv")
```

[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.

Exploratory analysis

1. Visualize, summarize, and describe the relationship between `Party` and `Tip`.

```
ggplot(tips, aes(x = Party, y = Tip)) +
  geom_point()
```



```
corr_coef <- tips %>%
  summarize(r = cor(Party, Tip)) %>%
  pull(r)
```

The relationship between Party and Tip is linear, moderately strong, and positive. The correlation coefficient between these variables is 0.79.

Modeling

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

2. Write the statistical model.

$$\hat{Tip} = \beta_0 + \beta_1 \times Party$$

or

$$Tip = \hat{\beta}_0 + \hat{\beta}_1 \times Party + \epsilon\epsilon = N(0, \sigma_\epsilon^2)$$

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.

```r
tips_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(Tip ~ Party, data = tips)

tidy(tips_fit) %>%
  kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.38	0.32	1.19	0.23
Party	1.96	0.12	16.55	0.00

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

Inference

Inference for the slope

6. The following code can be used to create a bootstrap distribution for the slope (and the intercept, though we'll focus primarily on the slope in our inference). Describe what each line of code does, supplemented by any visualizations that might help with your description.

```r
set.seed(1234)

boot_dist <- tips %>%
  specify(Tip ~ Party) %>%
  generate(reps = 100, type = "bootstrap") %>%
  fit()
```

7. Use the bootstrap distribution created in Exercise 6, `boot_dist`, to construct a 90% confidence interval for the slope using bootstrapping and the percentile method and interpret it in context of the data.

```
obs_fit <- tips %>%
  specify(Tip ~ Party) %>%
  fit()

get_confidence_interval(
  boot_dist,
  level = 0.90,
  type = "percentile",
  point_estimate = obs_fit
)
```

```
# A tibble: 2 x 3
  term      lower_ci upper_ci
  <chr>        <dbl>    <dbl>
1 intercept   -0.137     1.00
2 Party        1.69      2.21
```

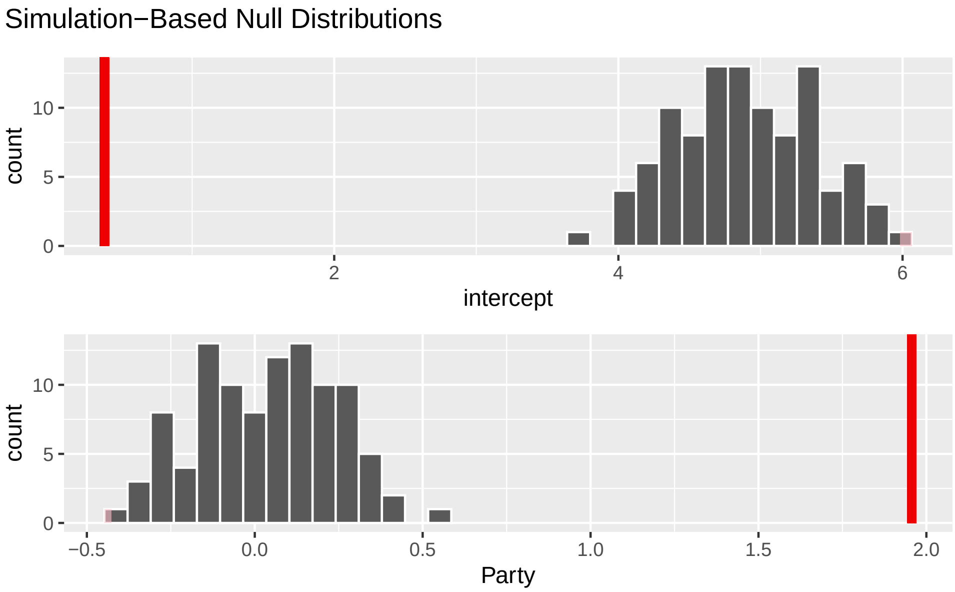
7. Conduct a hypothesis test at the equivalent significance level using permutation. State the hypotheses and the significance level you're using explicitly. Also include a visualization of the null distribution of the slope with the observed slope marked as a vertical line.

```
null_dist <- tips %>%
  specify(Tip ~ Party) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 100, type = "permute") %>%
  fit()

get_p_value(
  null_dist,
  obs_stat = obs_fit,
  direction = "two sided"
)
```

```
Warning: Please be cautious in reporting a p-value of 0. This result is an
approximation based on the number of `reps` chosen in the `generate()` step. See
`?get_p_value()` for more information.

Warning: Please be cautious in reporting a p-value of 0. This result is an
approximation based on the number of `reps` chosen in the `generate()` step. See
`?get_p_value()` for more information.
```

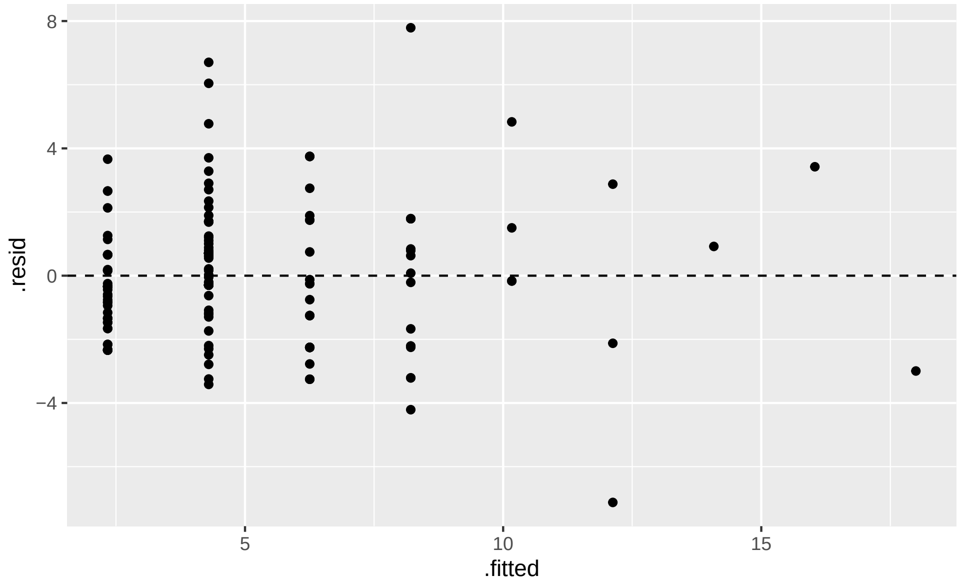
```
# A tibble: 2 x 2
  term      p_value
  <chr>       <dbl>
1 intercept       0
2 Party           0
```

```
visualize(null_dist) +
  shade_p_value(obs_stat = obs_fit, direction = "two sided")
```

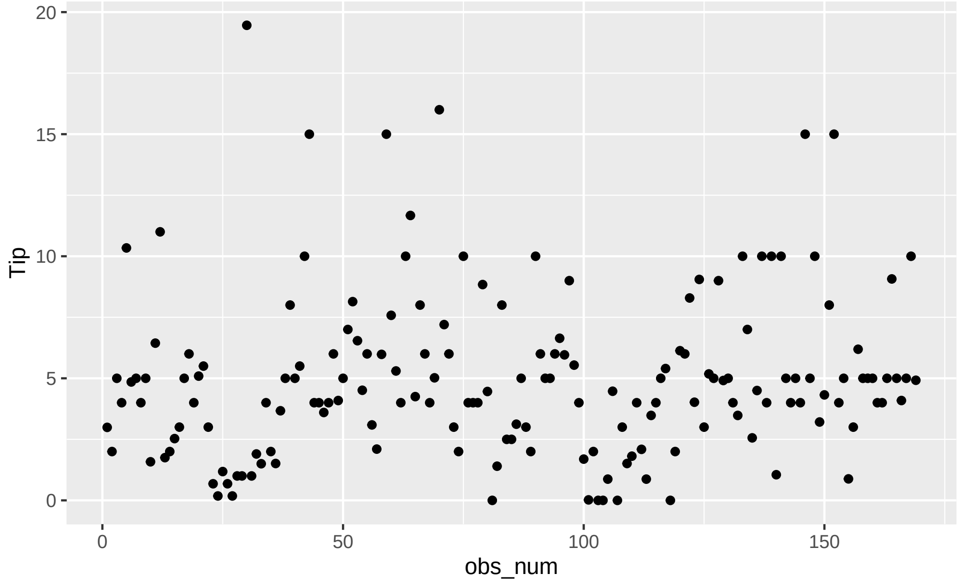


8. Check the relevant conditions for Exercises 7 and 8. Are there any violations in conditions that make you reconsider your inferential findings?

```
tips_aug <- augment(tips_fit$fit)

ggplot(tips_aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed")
```



```
tips_aug %>%
  mutate(obs_num = row_number()) %>%
  ggplot(aes(y = Tip, x = obs_num)) +
  geom_point()
```



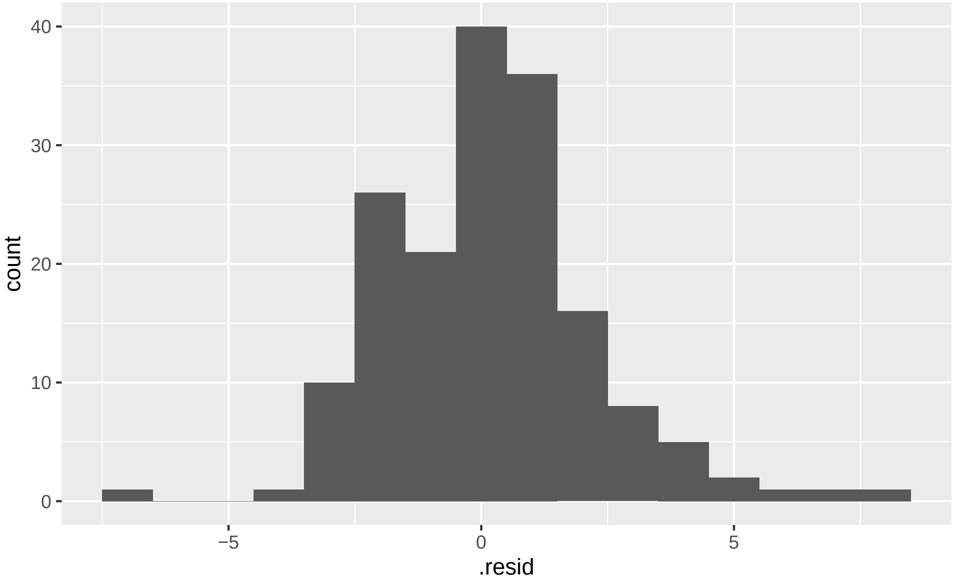
9. Now repeat Exercises 7 and 8 using approaches based on mathematical models.

```r
tidy(tips_fit, conf.int = TRUE, conf.level = 0.90)
```

```
# A tibble: 2 x 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)    0.383     0.321      1.19 2.34e- 1   -0.147     0.913
2 Party          1.96      0.118     16.6  4.77e-37    1.76      2.15
```

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

```r
ggplot(tips_aug, aes(x = .resid)) +
  geom_histogram(binwidth = 1)
```



Inference for a prediction

11. Based on your model, predict the tip for a party of 4.

```r
party_4 <- tibble(Party = 4)

predict(tips_fit, new_data = party_4)
```

```
# A tibble: 1 x 1
  .pred
  <dbl>
1  8.21
```

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.

```
predict(tips_fit, new_data = party_4, type = "conf_int")
```

```
# A tibble: 1 x 2
  .pred_lower .pred_upper
        <dbl>       <dbl>
1        7.71        8.71
```

```
predict(tips_fit, new_data = party_4, type = "pred_int")
```

```
# A tibble: 1 x 2
  .pred_lower .pred_upper
        <dbl>       <dbl>
1        4.07        12.4
```

Model diagnostics

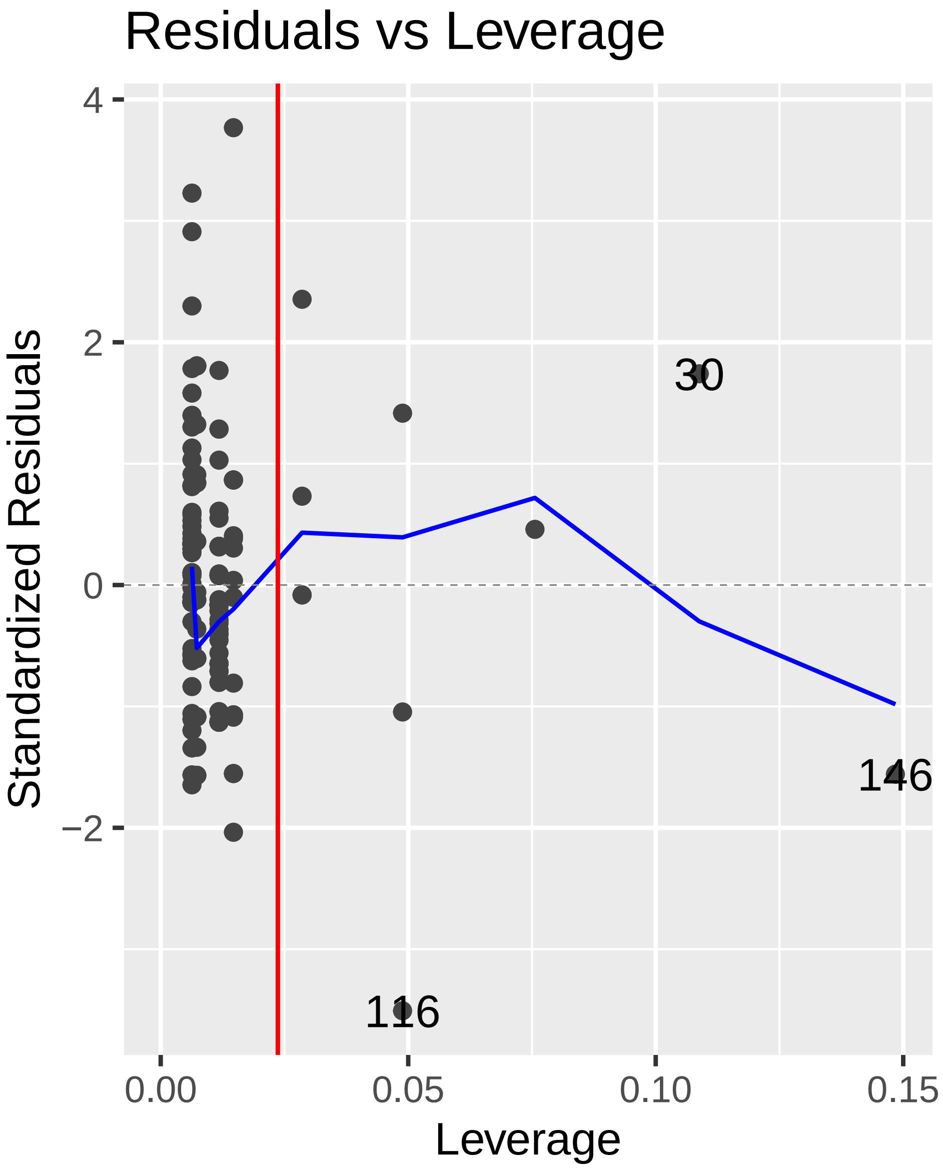
Leverage (Outliers in x direction)

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

```
leverage_threshold <- (2 * 2) / nrow(tips_aug)
```

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

```
autoplot(tips_fit$fit, which = 5) +
  geom_vline(xintercept = leverage_threshold, color = "red")
```



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

```
tips_aug %>%
  filter(.hat > leverage_threshold)
```

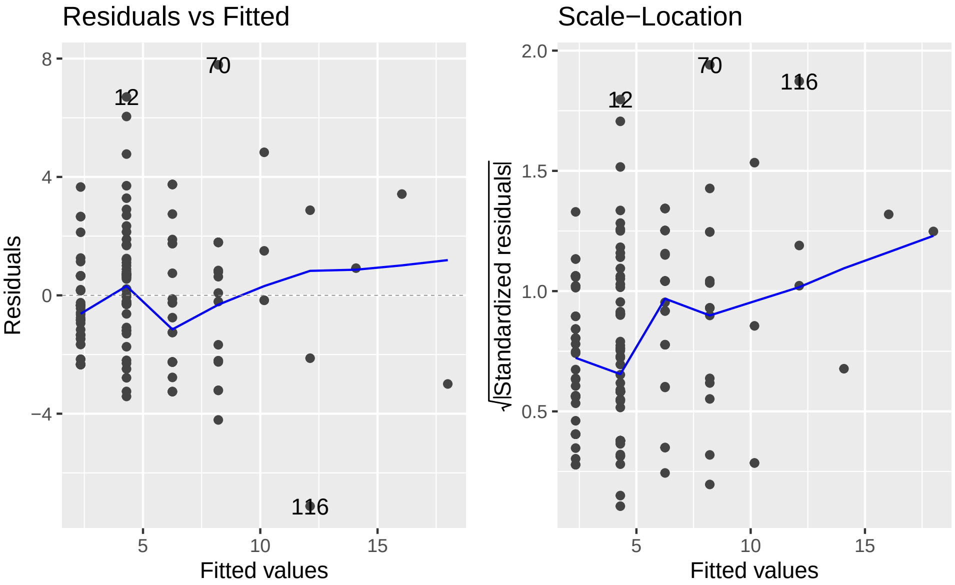
```
# A tibble: 10 x 8
     Tip Party .fitted .resid   .hat .sigma   .cooksd .std.resid
   <dbl> <dbl>   <dbl>  <dbl>  <dbl>  <dbl>     <dbl>      <dbl>
 1  19.5     8    16.0  3.42  0.109    2.07 0.185         1.74
 2  15       7    14.1  0.919 0.0756   2.09 0.00861       0.459
 3  15       5    10.2  4.83  0.0286   2.05 0.0815        2.35
 4  11.7     5    10.2  1.50  0.0286   2.09 0.00788       0.732
 5  10       6    12.1 -2.12  0.0489   2.08 0.0281       -1.05
 6   5       6    12.1 -7.12  0.0489   2.01 0.316        -3.51
 7  10       5    10.2 -0.167 0.0286   2.09 0.0000976    -0.0815
 8  10       5    10.2 -0.167 0.0286   2.09 0.0000976    -0.0815
 9  15       9    18.0 -2.99  0.148    2.07 0.212        -1.56
10  15       6    12.1  2.88  0.0489   2.08 0.0515        1.42
```

Identifying outliers (outliers in y direction)

17. Make a plot of the residuals vs. fitted values and a plot of the square root of the absolute value of standardized residuals vs. fitted (You can use `autoplot(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 $\sqrt{|\text{standardized residuals}|}$ vs. fitted to check conditions and model diagnostics?

```
autoplot(tips_fit$fit, which = c(1, 3))
```



18. Are there any observations that are outliers?

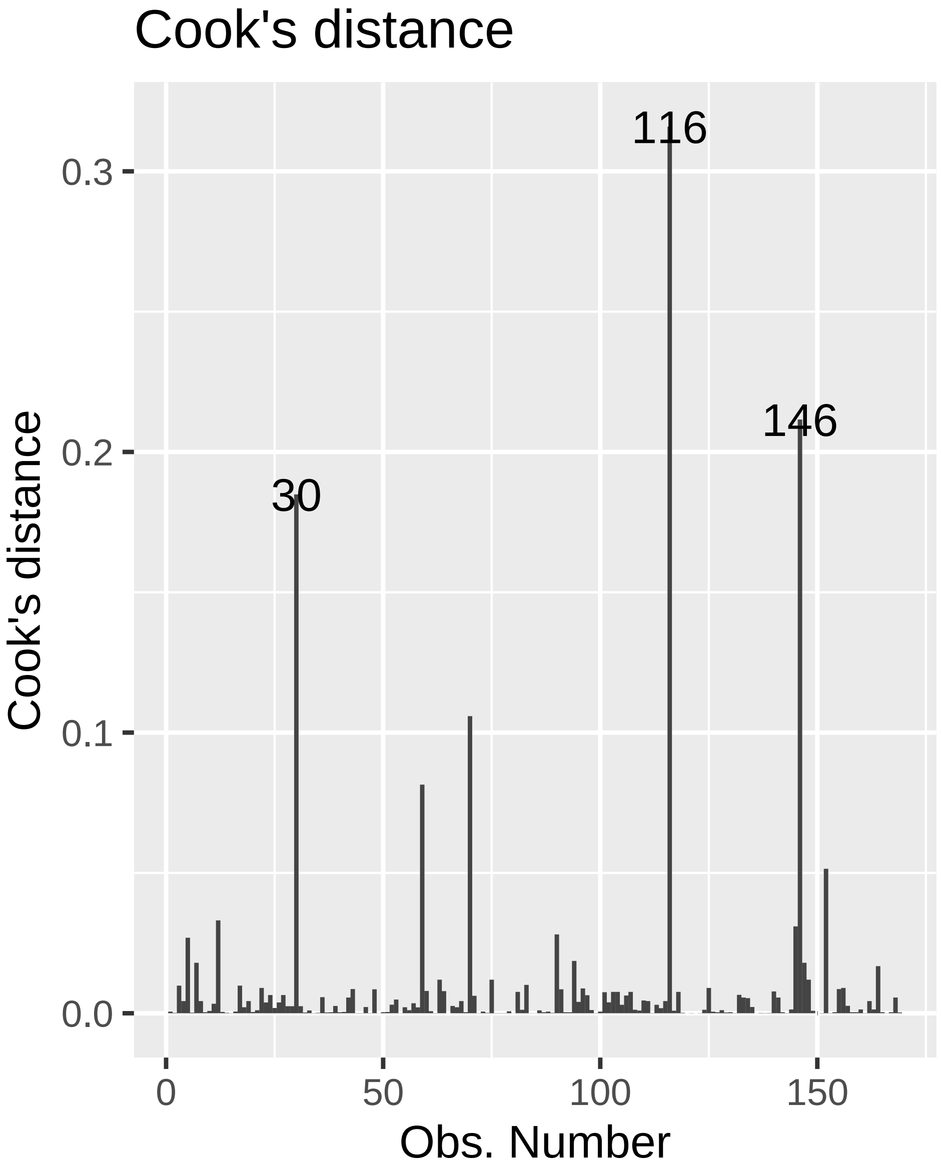
```
tips_aug %>%
  filter(.std.resid > 3 | .std.resid < -3)
```

```
# A tibble: 3 x 8
    Tip Party .fitted .resid    .hat .sigma .cooksd .std.resid
  <dbl> <dbl>   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>      <dbl>
1    11     2    4.30   6.70 0.00631   2.02  0.0331       3.23
2    16     4    8.21   7.79 0.0147    2.00  0.106        3.77
3     5     6   12.1   -7.12 0.0489    2.01  0.316       -3.51
```

Cook's distance

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

```
autoplot(tips_fit$fit, which = 4)
```

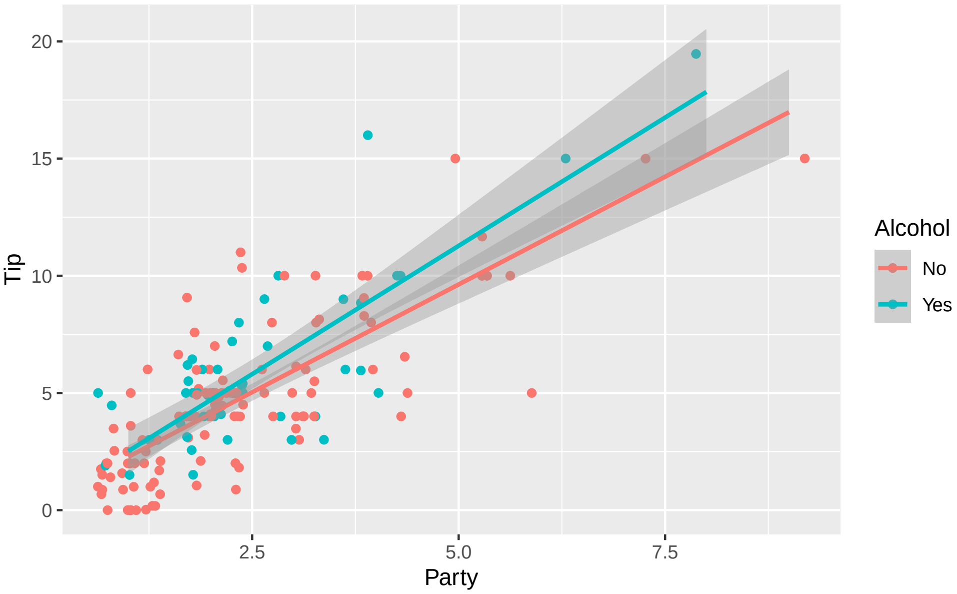


Adding another variable

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

```
ggplot(tips, aes(x = Party, y = Tip, color = Alcohol)) +
  geom_jitter() +
  geom_smooth(method = "lm")
```

```
`geom_smooth()` using formula 'y ~ x'
```



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

```r
tips_fit_2 <- linear_reg() %>%
  set_engine("lm") %>%
  fit(Tip ~ Party + Alcohol, data = tips)

tidy(tips_fit_2) %>%
  kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.22	0.33	0.69	0.49
Party	1.93	0.12	16.44	0.00
AlcoholYes	0.77	0.35	2.17	0.03

21. Interpret each of the slopes.
22. Does it make sense to interpret the intercept? Explain your reasoning.
23. According to this model, is the rate of change in tip amount the same for various sizes of parties regardless of alcohol consumption or are they different? Explain your reasoning.