AE 07: Exam 01 review

Restaurant tips

Sep 28, 2022

Packages

```
library(tidyverse)
library(tidymodels)
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.[1]

The variables we'll focus on for this analysis are

- `Tip`: amount of the tip
- `Party`: number of people in the party
- `Alcohol`: whether alcohol was purchased with meal

View the data set to see the remaining variables.

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

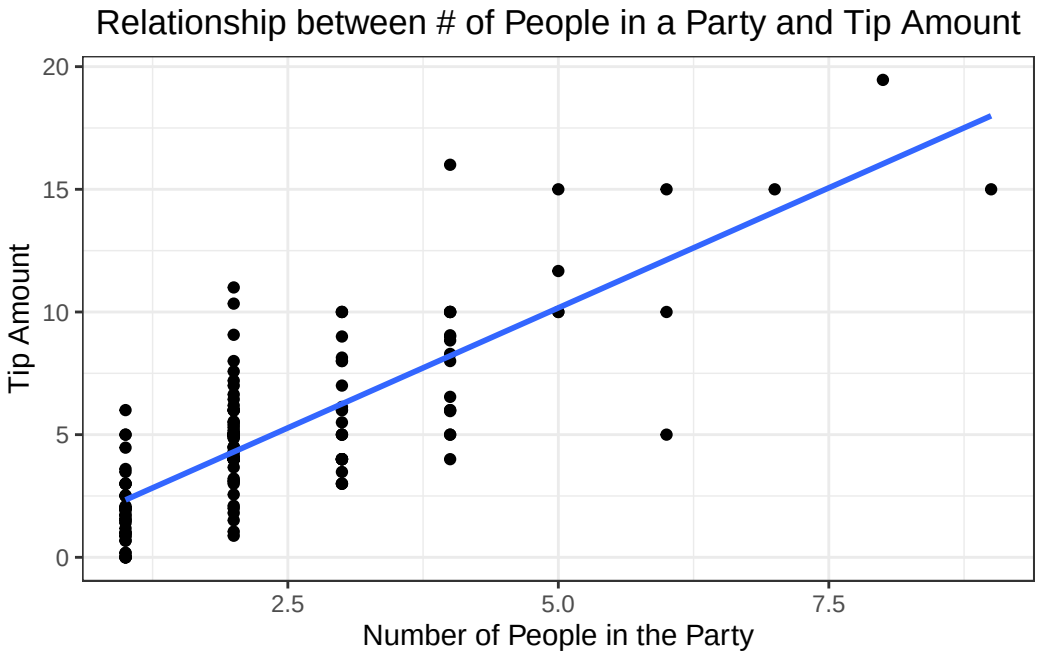
Exploratory analysis

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

[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.

```
ggplot(tips, aes(x = Party, y = Tip)) +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  theme_bw() +
  labs(x = "Number of People in the Party", y = "Tip Amount",
       title = "Relationship between # of People in a Party and Tip Amount") +
  theme(plot.title = element_text(hjust = 0.5))
```

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




```
cor(tips$Tip, tips$Party)
```

```
[1] 0.7882332
```

There is a clear linear relationship between the number of people in a party and the amount of the tip - as the number of people in the party increases, so too does the tip size. The relationship is fairly strong - with a correlation coefficient of .788, there is a relationship between the two variables.

Modeling (6 minutes)

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

2. Write the statistical model.

$$Tip = \beta_0 + \beta_1 \times Party + \epsilon$$

3. Fit the regression line and write the regression equation. Name the model `tips_fit` and neatly display the results with 3 digits and the 95% confidence interval for the coefficients.

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

tidy(tips_fit, conf.int = T, conf.level = .95) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.383	0.321	1.195	0.234	-0.250	1.016
Party	1.957	0.118	16.553	0.000	1.723	2.190

$$\hat{Tip} = 0.383 + 1.957 \times Party$$

4. Interpret the slope.

For every 1 person increase in our party size, we expect our tip size to increase by 1.957 dollars, on average.

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

No it does not - you cannot have zero people in a party.

Inference

Inference for the slope (8 minutes)

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.

```
set.seed(1234) #Sets the seed for our random number generator, so that we
#replicate the same findings every single time we run our code

boot_dist <- tips |> #Create a new object "boot_dist", and assign it to tips
  specify(Tip ~ Party) |> #Specify which two variables we should find a
  #relationship between, for the purpose of generating bootstrap samples
  generate(reps = 100, type = "bootstrap") |> #Generate 100 bootstrap samples
  #of the relationship between Tip and Party
  fit() #Gives us our bootstrapped distribution in a tidy output, with only
#one slope and intercept per rep
```

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,
  point_estimate = obs_fit,
  level = .9,
  type = "percentile"
)
```

```
# 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
```

We are 90% confident that the true population slope of the relationship between party size and tip value lies on the interval (1.69, 2.21).

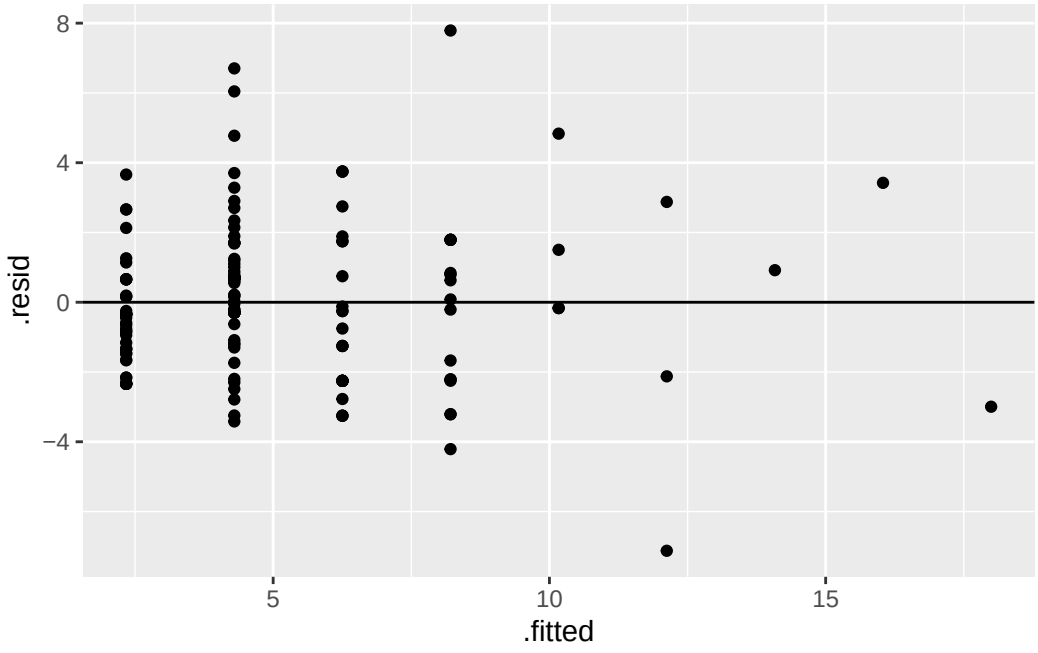
8. Conduct a hypothesis test at the equivalent significance level using permutation with 100 reps. State the hypotheses and the significance level you're using explicitly. Also include a visualization of the null distribution of the slope with the observed slope marked as a vertical line.

```
set.seed(1234)
# add your code here
```

9. 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)
```



Simulation inference: linearity & independence.

Independence: satisfied. We can reasonably assume what happened at one table does not provide information about what happened at another table.

Linearity: satisfied. There is no pattern in the plot of residuals versus fitted values.

10. Now repeat Exercises 7 and 8 using approaches based on mathematical models. You can reference output from previous exercises and/or write new code as needed.
11. Check the relevant conditions for Exercise 10. Are there any violations in conditions that make you reconsider your inferential findings? You can reference previous graphs / conditions and add any new code as needed.

```
# add your code here
```

Linearity:

Independence:

Constant variance (critical for mathematical models):

Normality (relaxed in n $>$ 30):

Inference for a prediction (5 minutes)

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

```
# add your code here
```

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

```
# add your code here
```

Multiple linear regression (6 minutes)

15. Make a plot to visualize the relationship between `Party` and `Tip` with the points colored by `Alcohol`. Describe any patterns that emerge.

```
# add your code here
```

16. Fit a multiple linear regression model predicting `Tip` from `Party` and `Alcohol`. Display the results with `kable()` and three digits.

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

tidy(multiModel) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.225	0.325	0.691	0.490
Party	1.932	0.118	16.437	0.000
AlcoholYes	0.765	0.353	2.166	0.032

17. Interpret the coefficients of `Party` and `Alcohol`.

18. 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.

Important

To submit the AE:

- Render the document to produce the PDF with all of your work from today's class.
- Push all your work to your `ae-07-` repo on GitHub. (You do not submit AEs on Gradescope).