STA 310: Homework 1

Luke Flyer

Instructions

- Write all narrative using full sentences. Write all interpretations and conclusions in the context of the data.
- Be sure all analysis code is displayed in the rendered pdf.
- If you are fitting a model, display the model output in a neatly formatted table. (The `tidy` and `kable` functions can help!)
- If you are creating a plot, use clear and informative labels and titles.
- Render and back up your work reguarly, such as using Github.
- When you're done, we should be able to render the final version of the Rmd document to fully reproduce your pdf.
- Upload your pdf to Gradescope. Upload your Rmd, pdf (and any data) to Canvas.

Exercises

Exercises 1 - 4 are adapted from exercises in Section 1.8 of @roback2021beyond.

Exercise 1

Consider the following scenario:

> Researchers record the number of cricket chirps per minute and temperature during that time. They use linear regression to investigate whether the number of chirps varies with temperature.

a. Identify the response and predictor variable.

The response variable is the number of chirps per minute, and they predictor variable is temperature.

b. Write the complete specification of the statistical model.

$Y_i = \beta_0 + \beta_1(Temperature_i) + \epsilon_i$

In this case, β_0 represents the true intercept—the expected number of cricket chirps per minute at temperature = 0. β_1 represents the true slope—the expected increase in cricket chirps per minute for every unit increase in temperature, assuming the rate of increase is linear. ϵ_i represent the deviations of the actual chirps per minute from the expected number predicted by the model. Y_i represents the real observations of cricket chirps per minute.

c. Write the assumptions for linear regression in the context of the problem.

Linearity: The mean number of cricket chirps per minute is linearly related to temperature.

Independence: Each observation of cricket chirps per minute and temperature is independent and not related to each other.

Normality: The cricket chirps per minute are normally distributed at any given temperature.

Equal variance: Variability of cricket chirps per minute is the same at all temperatures.

Exercise 2

Consider the following scenario:

> A randomized clinical trial investigated postnatal depression and the use of an estrogen patch. Patients were randomly assigned to either use the patch or not. Depression scores were recorded on 6 different visits.

a. Identify the response and predictor variables.

   The response variable is depression score, and the predictor variable is estrogen patch vs. control group.

b. Identify which model assumption(s) are violated. Briefly explain your choice.

   Independence seems to be most violated since depression scores were recorded on 6 different visits, meaning that there are 6 observations for each patient. This would mean that knowing about one of those observations could provide information about another.

Exercise 3

Use the Kentucky Derby case study in Chapter 1 of *Beyond Multiple Linear Regression.*

a. Consider Equation (1.3) in Section 1.6.3. Show why we have to be sure to say "holding year constant", "after adjusting for year", or an equivalent statement, when interpreting β_2.

   We have to say "holding year constant" because the year is another predictor in this linear model. Therefore, β_2 is the expected difference in winning speeds between slow and fast conditions after accounting for year.

b. Briefly explain why there is no error (random variation) term ϵ_i in Equation (1.4) in Section 1.6.6?

   This is because this equation gives us "estimated winning speeds" which does not have to include this error term. The actual observations from the model will have error, but the estimation equation does not.

Exercise 4

The data set `kingCountyHouses.csv` in the `data` folder contains data on over 20,000 houses sold in King County, Washington (@kingcounty).

We will use the following variables:

- `price` = selling price of the house
- `sqft` = interior square footage

See Section 1.8 of Beyond Multiple Linear Regression for the full list of variables.

a. Fit a linear regression model with `price` as the response variable and `sqft` as the predictor variable (Model 1). Interpret the slope coefficient in terms of the expected change in price when `sqft` increases by 100.

```
library(tidyverse)
library(tidymodels)
library(knitr)
houses <- read_csv("/home/guest/310/homeworks/data/kingCountyHouses.csv")

linear_reg() |>
  set_engine("lm") |>
  fit(price ~ sqft, data = houses) |>
  tidy() |>
  kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-43580.7431	4402.689690	-9.898663	0
sqft	280.6236	1.936399	144.920356	0

For every increase of 100 square feet, the price of a house is expected to increase by $28,062.

b. Fit Model 2, where `logprice` (the natural log of price) is now the response variable and `sqft` is still the predictor variable. How is the `logprice` expected to change when `sqft` increases by 100?

```
linear_reg() |>
  set_engine("lm") |>
  fit(log(price) ~ sqft, data = houses) |>
  tidy() |>
  kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	12.2184641	0.0063741	1916.8830	0
sqft	0.0003987	0.0000028	142.2326	0

For every increase of 100 square feet, the log(price) of a house is expected to increase by .0399 log(dollars).

c. Recall that $log(a) - log(b) = log(\frac{a}{b})$. Use this to derive how the `price` is expected to change when `sqft` increases by 100 based on Model 2.

For every increase of 100 square feet, the price of a house is expected to be multiplied by a factor of $e^{0.0399}$. This is because $0.0399 = log(\frac{a}{b})$, and we then need to exponentiate both sides to solve for the price ratio of a house with 100 more square feet to the price of the house with 100 fewer.

d. Fit Model 3, where `price` and `logsqft` (the natural log of sqft) are the response and predictor variables, respectively. How does the price expected to change when sqft increases by 10%? *As a hint, this is the same as multiplying sqft by 1.10.*

```
linear_reg() |>
  set_engine("lm") |>
  fit(price ~ log(sqft), data = houses) |>
  tidy() |>
  kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-3451377.1	35169.348	-98.13594	0
log(sqft)	528647.5	4650.631	113.67221	0

```
log(1.1) * 528647.5
```

```
## [1] 50385.49
```

For a 10% increase in square feet (multiplicative increase of 1.1), we expect the price of a house to increase by $50385.49.

Click here for notes on interpreting model effects for log-transformed response and/or predictor variables.

Exercise 5

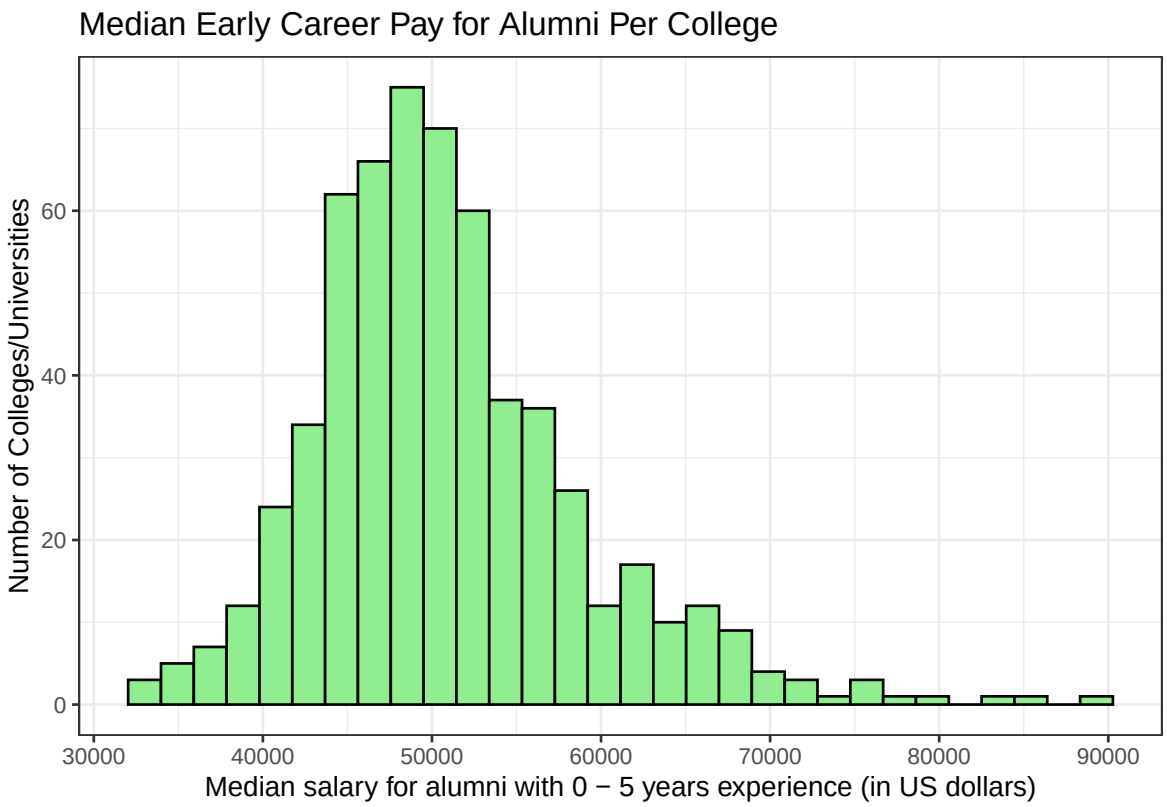
The goal of this analysis is to use characteristics of 593 colleges and universities in the United States to understand variability in the early career pay, defined as the median salary for alumni with 0 - 5 years of experience. The data was obtained from TidyTuesday College tuition, diversity, and pay, and was originaly collected from the PayScale College Salary Report.

The data set is located in `college-data.csv` in the `data` folder. We will focus on the following variables:

variable	class	description
name	character	Name of school
state_name	character	state name
type	character	Public or private
early_career_pay	double	Median salary for alumni with 0 - 5 years experience (in US dollars)
stem_percent	double	Percent of degrees awarded in science, technology, engineering, or math subjects
out_of_state_total	double	Total cost for in-state residents in USD (sum of room & board + out of state tuition)

a. Visualize the distribution of the response variable `early_career_pay`. Write 1 - 2 observations from the plot.

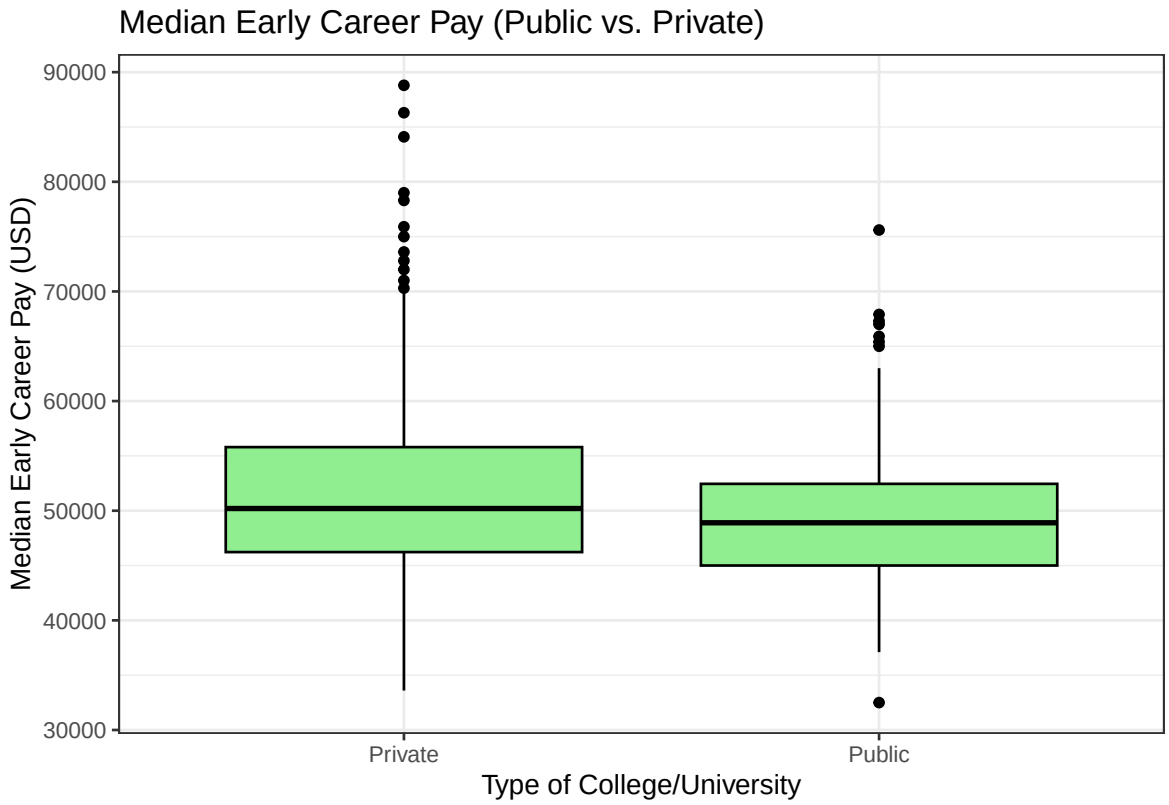
```
college <- read_csv("/home/guest/310/homeworks/data/college-data.csv")

college |>
  ggplot(aes(x = early_career_pay)) +
  geom_histogram(color = "black", fill = "light green") +
  theme_bw() +
  labs(
    x = "Median salary for alumni with 0 - 5 years experience (in US dollars)",
    y = "Number of Colleges/Universities",
    title = "Median Early Career Pay for Alumni Per College")
```



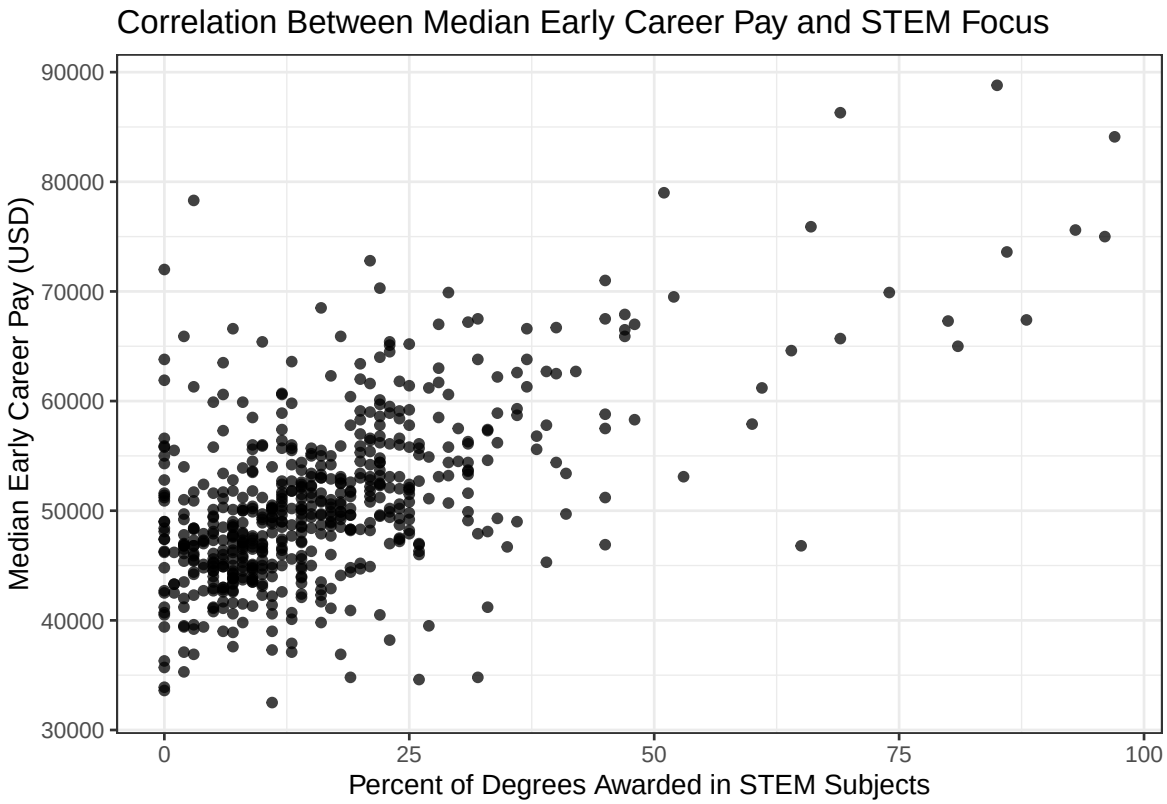
Most of the data seems to be captured between the \$40,000 and \$60,000 marks, with a median somewhere around \$50,000. Additionally, the data is skewed left, with a longer right tail containing some schools whose students have very large early salaries.

b. Visualize the relationship between (i) `early_career_pay` and `type` and (ii) `early_career_pay` and `stem_percent`. Write an observation from each plot.

```
college |>
  ggplot(aes(y = early_career_pay, x = type)) +
  geom_boxplot(color = "black", fill = "light green") +
  theme_bw() +
  labs(
    x = "Type of College/University",
    y = "Median Early Career Pay (USD)",
    title = "Median Early Career Pay (Public vs. Private)"
    )
```




```
college |>
ggplot(aes(x = stem_percent, y = early_career_pay)) +
geom_point(alpha = 0.75) +
theme_bw() +
labs(x = "Percent of Degrees Awarded in STEM Subjects",
y = "Median Early Career Pay (USD)",
title = "Correlation Between Median Early Career Pay and STEM Focus")
```



From the first plot, we can see that the distributions between private and public schools seem to be pretty similar, one difference being that private schools have a higher cap in early career pay, with some schools landing alumni in the \$80K-\$90K range. From the second plot, we can see a relatively clear positive correlation between early career pay and the percent of STEM degrees awarded, which makes sense given the generally higher pay of STEM careers.

c. Below is the specification of the statistical model for this analysis. Fit the model and neatly display the results using 3 digits. Display the 95% confidence interval for the coefficients.

$$early\_career\_pay_i = \beta_0 + \beta_1\ out\_of\_state\_total_i + \beta_2\ type \tag{1}$$
$$+ \beta_3\ stem\_percent_i + \beta_4\ type * stem\_percent_i \tag{2}$$
$$+ \epsilon_i, \quad \text{where } \epsilon_i \sim N(0, \sigma^2) \tag{3}$$

```
linear_reg() |>
  set_engine("lm") |>
  fit(early_career_pay ~ out_of_state_total + type + stem_percent + type * stem_percent,
data = college) |>
  tidy(conf.int = TRUE) |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	36217.704	850.222	42.598	0.000	34547.862	37887.546
out_of_state_total	0.253	0.018	13.692	0.000	0.217	0.289
typePublic	1185.020	768.752	1.541	0.124	-324.813	2694.853

term	estimate	std.error	statistic	p.value	conf.low	conf.high
stem_percent	214.306	19.300	11.104	0.000	176.402	252.211
typePublic:stem_percent	49.538	33.875	1.462	0.144	-16.992	116.069

d. How many degrees of freedom are there in the estimate of the regression standard error σ?

There are 593 - 5 = 588 degrees of freedom in the estimate of the regression standard error σ because there are 5 coefficients in the model, and that number is subtracted from the sample size.

e. What is the 95% confidence interval for the amount in which the intercept for public institutions differs from private institutions?

We are 95% confident that the true mean difference in median early career pay between public and private institutions is between $-324.813 and $2,694.853.

Exercise 6

Use the analysis from the previous exercise to write a paragraph (~ 4 - 5 sentences) describing the differences in early career pay based on the institution characteristics. *The summary should be consistent with the results from the previous exercise, comprehensive, answers the primary analysis question, and tells a cohesive story (e.g., a list of interpretations will not receive full credit).*

During this exploratory data analysis, we first examined the differences in early career pay between private and public schools. The distributions looked very similar, with the only difference being a longer right tail for private schools reaching the $90,000 mark while public schools only reached around $70,000. In the regression analysis, the 95% confidence interval for the difference in early career pay between public and private institutions (holding all other variables constant) included 0, meaning that the effect of school type (public vs. private) on early career pay is not significant (holding all other variables constant). A scatterplot comparing early career pay and STEM degree percentage appeared to show a positive correlation, and the regression analysis bore this out. For private universities, for every 1 percent increase in STEM degree percentage, we expect a $214.306 increase in early career pay (holding all other variables constant). For public universities, for every 1 percent increase in STEM degree percentage, we expect a $214.306 + $49.538 = $263.844 increase in early career pay (holding all other variables constant). The p-value associated with the STEM degree variable was less than 0.05, so we can say that the effect of STEM degree percentage on early career pay is significant (holding all other variables constant). However, the p-value associated with the interaction effect between STEM degrees and school type (public vs. private) was not less than 0.05. Therefore, we can say that there is insufficient evidence to suggest that the effect of STEM degree percentage on early career pay depends on school type (holding all other variables constant). Finally, we also found in the regression model that for every $1 increase in out of state total cost, we would expect an increase of $0.253 in early career pay (holding all other variables constant). This effect was significant, suggesting a positive relationship between out of state cost and early career pay (holding all other variables constant). Overall, these results help us understand some possible explanations for variability in early career pay among alumni.

Grading

Total	**50**
Ex 1	8

Total	**50**
Ex 2	4
Ex 3	7
Ex 4	12
Ex 5	12
Ex 6	4
Workflow & formatting	3

The "Workflow & formatting" grade is to based on the organization of the assignment write up along with the reproducible workflow. This includes having an organized write up with neat and readable headers, code, and narrative, including properly rendered mathematical notation. It also includes having a reproducible Rmd/Quarto document that can be rendered to reproduce the submitted PDF.