STA 310: Homework 1

Pascal Bell

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 cricket chirps per minute, and predictor variable is the temperature.

b. Write the complete specification of the statistical model.

   The complete specification of the statistical model is the following:

   $y_{ij} = x_i^T * B + e_{ij}$

   Where y_{ij} is the temperature of the jth observation of the ith group and x_i is the mean cricket chirps for the ith group. B represents the estimated change in cricket chirps per minute for a one unit increase in temperature. The e_{ij} term represents the error, or the difference between the predicted chirps per minute, and the actual chirps per minute for a given temperature.

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

   1. Linearity: Linear relationship between mean cricket chirps per minute and temperature
   2. Independence: Each unique observation of temperature and cricket chirps are independent and not related to each other
   3. Normality: The distribution of cricket chirps follows a normal distribution at each temperature
   4. Equal variance: Variability of cricket chirps is equal for 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 the depression score from a patient, and the predictor variable is whether or not they use an estrogen patch.

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

   The linearity assumption is not met since the outcome is a categorical variable. Therefore, there cannot be a linear relationship between use of estrogen patch and the depression score from a patient as the outcome must be continuous for linear regression.

   The independence assumption is also violated since it is not specified if the patients were randomly selected or if they were volunteers. If the were not randomly selected, there could be confounding variables that make the outcomes not independent.

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” when interpreting β_2 because if year is not held constant, and the response is compared when year changes, we would be adding an additional change to the response. To interpret $beta_2$ we want to examine the change in the response by a change in only the predictor Fast; if we do not specify that year was held constant in this interpretation, we do not clarify that the β_2 is the change in the response when all other predictors are held constant.

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

   The equation provided represents the predicted regression model that is used to estimate the response. Therefore there is no error term since the predicted value falls on the line of best fit provided by the equation; the error term is 0.

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.

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

```r
houses <- read_csv("../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

The price of a house is expected to increase by $28,062 on average when `sqft` increases by 100 ft.

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?

```r
houses <- houses |>
  mutate(logprice = log(price))

linear_reg() |>
  set_engine("lm") |>
  fit(logprice ~ 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

`logprice` of a house is expected on average to increase by .0399 when `sqft` increases by 100

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.

`price` of a house is expected on average to be increased by a factor of $e^{0.0399}$ when `sqft` increases by 100 ft. This is because we know that $0.0399 = log(\frac{a}{b})$ where a and b represent the price when `sqft` is 100 and 0 ft respectively. To solve for the ratio of when price when `sqft` is 100 to 0, we simply exponentiate both sides.

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.*

```r
houses <- houses |>
  mutate(logsqft = log(sqft))

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

```
## # A tibble: 2 x 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) -3451377.    35169.     -98.1       0
## 2 logsqft       528647.     4651.     114.        0
```

The slope coefficient for `logsqft` is interpreted as in terms of the change in the mean of `price` when `sqft` is multiplied by e. So for a 10% change in `sqft` we would multiply the slope coefficient of `logsqft` by $528647.5 * log(1.1)$. Therefore, for a 10% change in `sqft` we would expect `price` to be increased on average by $50,385.49.

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

Exercise 5

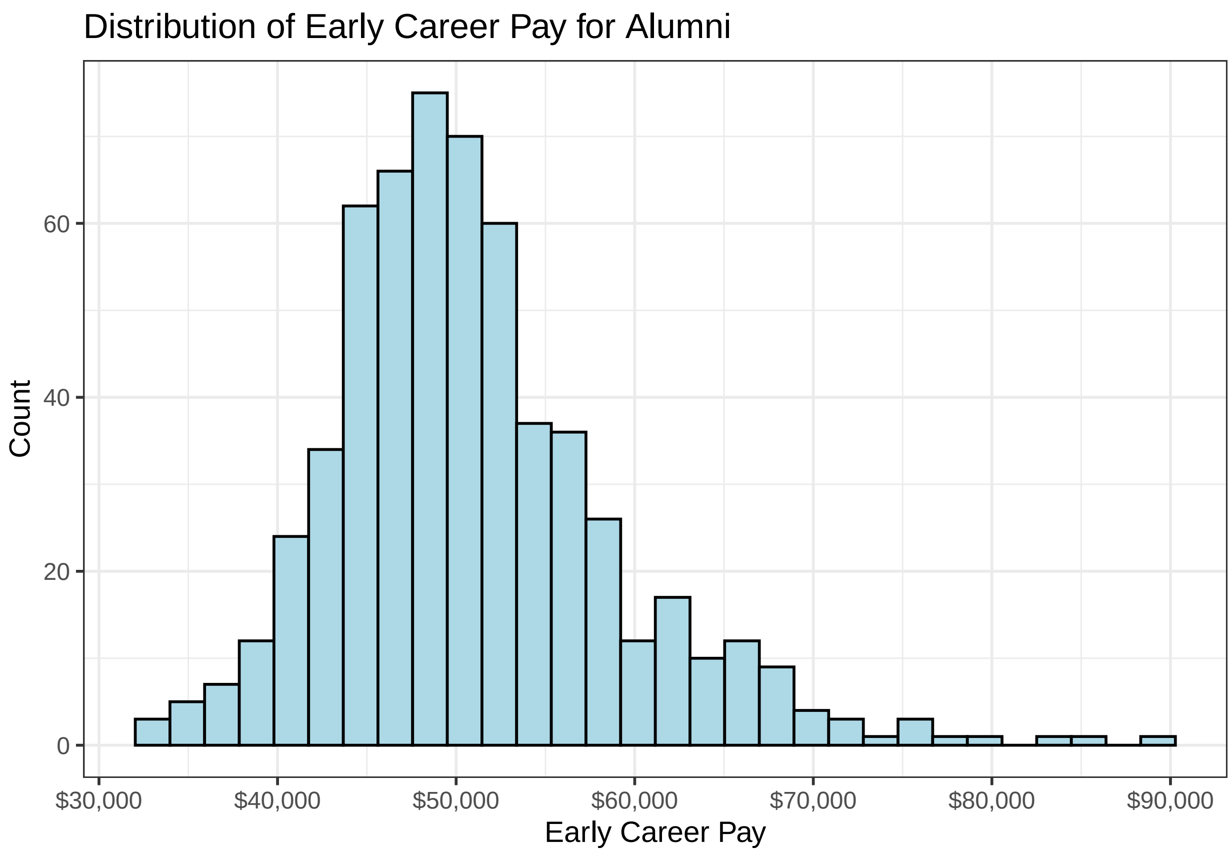
The goal of this analysis is to use characteristics of 593 colleges and universities in the United States to understand variability in the early career pay, defined as the median salary for alumni with 0 - 5 years of experience. The data was obtained from TidyTuesday College tuition, diversity, and pay,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

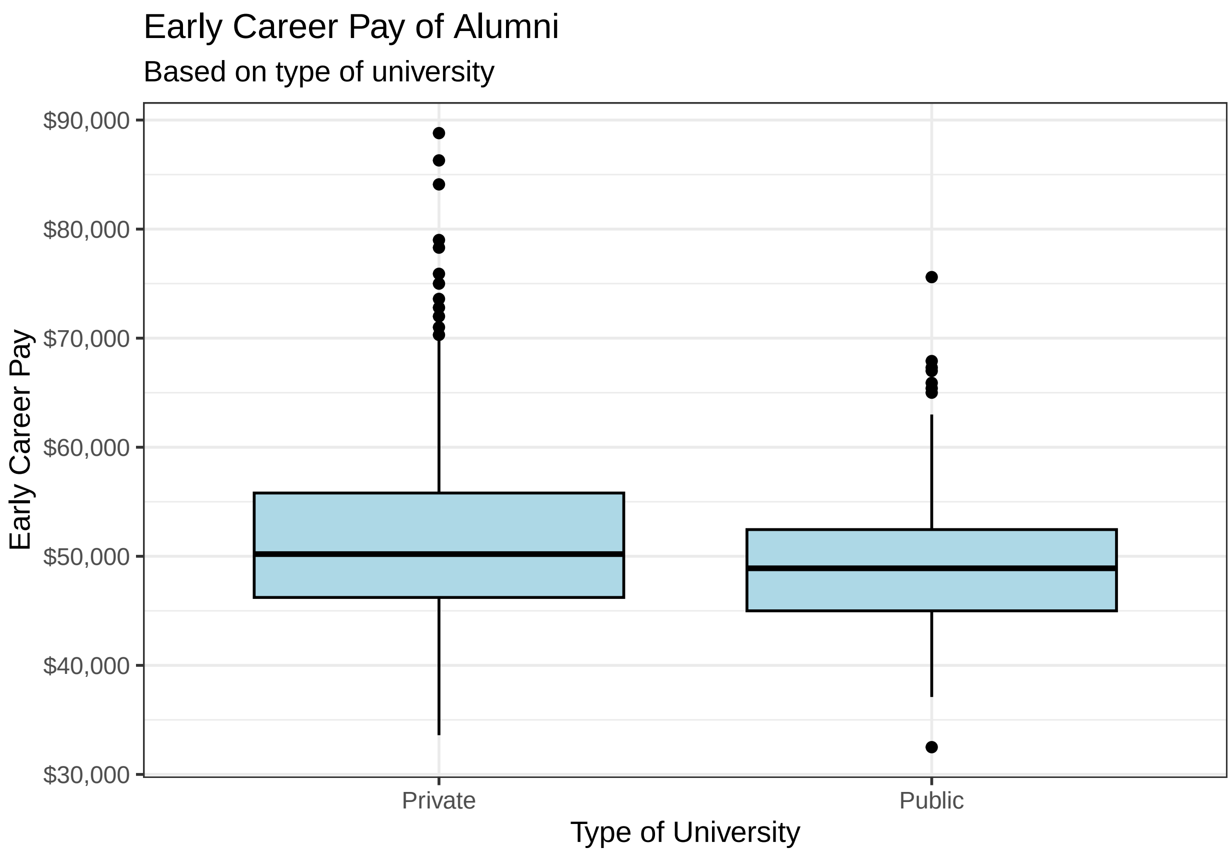
```r
salaries <- read_csv("../data/college-data.csv")
salaries |>
  ggplot(aes(x = early_career_pay)) +
  geom_histogram(color = "black", fill = "lightblue") +
  theme_bw() +
  labs(x = "Early Career Pay",
       y = "Count",
       title = "Distribution of Early Career Pay for Alumni") +
  scale_x_continuous(labels = scales::dollar_format())
```



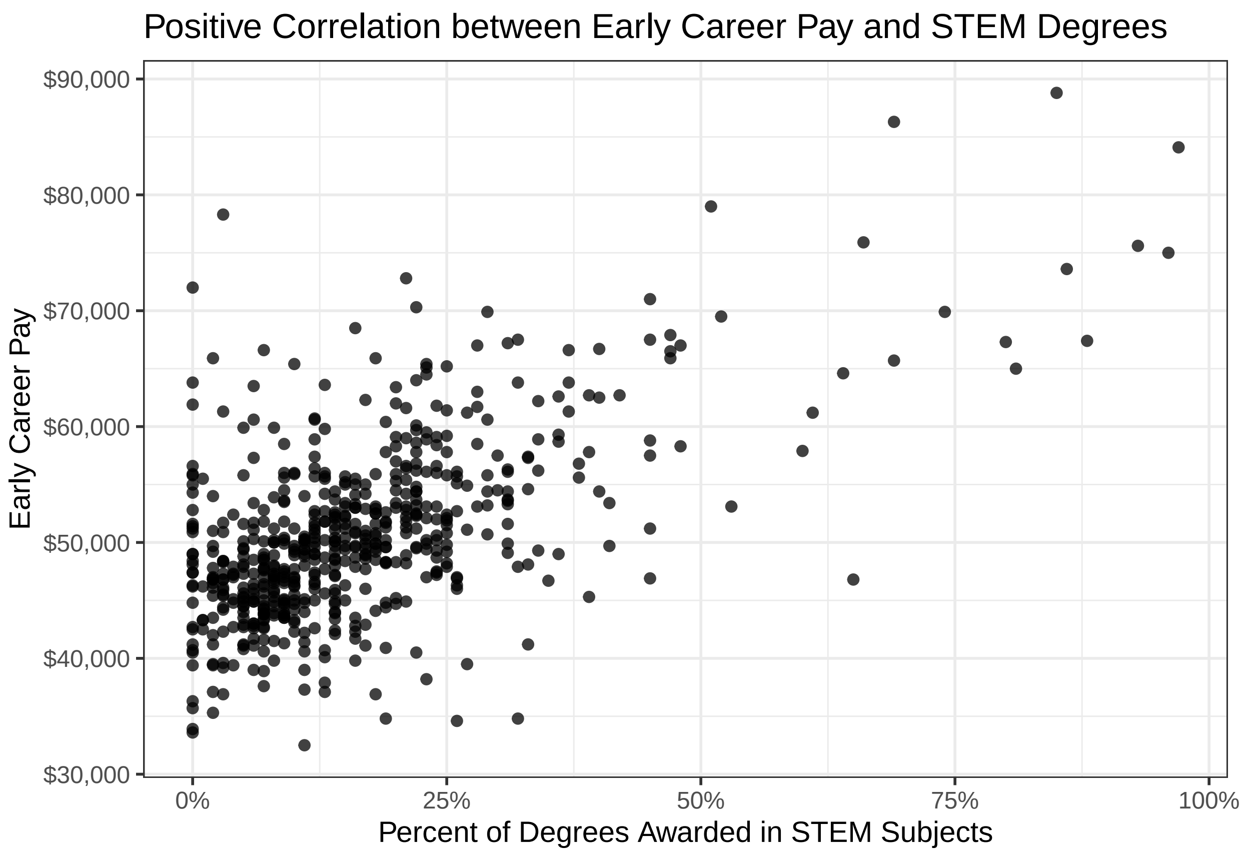
1. One observation from this plot is that most of the data is centered within the $40,000-$60,000 range
2. Another observation from this plot is that the data seems slightly skewed right with a tail on the right side, where there are a handful of outliers.

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.

```
salaries |>
  ggplot(aes(x = type, y = early_career_pay)) +
  geom_boxplot(color = "black", fill = "lightblue") +
  theme_bw() +
  labs(x = "Type of University",
       y = "Early Career Pay",
       title = "Early Career Pay of Alumni",
       subtitle = "Based on type of university") +
  scale_y_continuous(labels = scales::dollar_format())
```



```
salaries |>
  mutate(stem_percent = stem_percent / 100) |>
  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 = "Early Career Pay",
       title = "Positive Correlation between Early Career Pay and STEM Degrees") +
  scale_y_continuous(labels = scales::dollar_format()) +
  scale_x_continuous(labels = scales::percent_format())
```



An observation from the first plot is that private universities have many more outliers with schools where the median early career pay is over $70,000, in comparison to public universities. It seems like private universities have a higher cap for early career pay.

An observation from the second plot is that there seems to be a positive correlation between the percent of degrees awarded in STEM subjects and early career pay for alumni, where schools with more stem degrees tend to have a higher median early career pay.

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.

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

```r
linear_reg() |>
  set_engine("lm") |>
  fit(early_career_pay ~ out_of_state_total + type + stem_percent + type * stem_percent,
      data = salaries) |>
  tidy(conf.int = TRUE) |>
  mutate(across(where(is.numeric), ~ round(.x, 3))) |>
  kable()
```

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
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?

```
salaries |>
  nrow()
```

```
## [1] 593
```

There are 588 degrees of freedom for this estimate of the regression standard error. There are 593 total rows in the dataset and 5 coefficients in the model (including the intercept), so there are 588 total degrees of freedom.

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).*

Based on exploratory data analysis, it seemed like public and private universities lead to similar early career pay for alumni; however, there are more higher outliers with some private schools where alumni have a median early career pay of over $70,000. It also seems like generally the more stem degrees a school has, the more early career pay this results in. These results are also quantified by our regression model which shows that a one percent increase in STEM degrees results in an expected increase of $214.306 for early career pay, for private institutions, and when all other variables are held constant. For a public university with zero percent stem degrees, we expect an increase of $1,185.02 in early career pay compared to private universities, while holding all other variables constant. If a university is public we expect an increase of $49.538 in early career salary every one percent increase in stem degrees, when all other variables are held constant. It is important to note that the confidence interval for the true mean difference in median early career pay between public and private institutions does contain zero, and its p-value is not statistically significant, so we can not conclude that there is a difference in the true mean difference in early career pay between public and private institutions. Our regression model also found that for every $1 increase in out of state total cost, we would expect an increase of $0.253 in early career pay. These results help us understand the variability in early career pay for college alumni.

Grading

Total	**50**
Ex 1	8
Ex 2	4

Total	**50**
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.