STA 310: Homework 1

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 temperature is the predictor variable. The number of cricket chirps per minute is the response variable.

b. Write the complete specification of the statistical model.

Temperature $= \beta_0 + \beta_1 \cdot$ Chirps per Minute $+ \epsilon$

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

Linearity: There is a linear relationship between the number of chirps per minute and temperature is linear.

Independence: The residuals of chirps and corresponding temperatures are independent of each other. There is no connection between how far any two points lie above or below the regression line.

Normality: The number of chirps per minute follows a normal distribution at each level of the temperature.

Equal Variance: The variability of the number of chirps per minute is equal for all levels of the temperature.

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 variables are the depression scores. The predictor variable is the use of an estrogen patch.

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

The assumption of Independence is violated. Depression scores are measured repeatedly over six visits for each patient, meaning the scores will be correlated to each patient.

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.

In the model given in Equation 1.3, we must provide the qualifiers when interpreting β_2 because in multiple linear regression when you add covariables like β_1 you must hold them constant to assess the effect of other variables. In the model provided, to make judgements on the predictor variable fast, you must hold year constant.

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

In Equation 1.4, the model is finding estimated winning speeds, $\hat{Y}_i$. This is different from finding the observed winning speeds. In fact, the error term accounts for the difference between the observed and predicted value. Thus, there is no error term needed in the equation for estimated values.

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.

Loading Packages and Data:

```
library(tidyverse)
library(tidymodels)
library(skimr)
houses <- read_csv("~/STA310/kingCountyHouses.csv")
```

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.

```
model1 <- linear_reg() |>
    set_engine("lm") |>
    fit(price ~ sqft, data = houses) |>
    tidy()

model1
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -43581.   4403.       -9.90 4.72e-23
## 2 sqft            281.      1.94    145.   0
```

The expected change in selling price when the square footage increases by 100 square feet is an increase of $28,062.36. This figure is β_1 from Model 1 multiplied by 100.

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?

```
model2 <- linear_reg() |>
  set_engine("lm") |>
  fit(log(price) ~ sqft, data = houses)

tidy(model2)
```

```
## # A tibble: 2 x 5
##   term         estimate  std.error statistic p.value
##   <chr>           <dbl>      <dbl>     <dbl>   <dbl>
## 1 (Intercept) 12.2      0.00637        1917.       0
## 2 sqft         0.000399 0.00000280      142.       0
```

The expected change in 'logprice' when the square footage increases by 100 square feet is .03987465. This figure is β_1 from Model 2 multiplied 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.

```
model2_coeff <- tidy(model2)
sqft_coef <- model2_coeff$estimate[model2_coeff$term == "sqft"]
sqft_coef
```

```
## [1] 0.0003987465
```

```
percent_change <- (exp(sqft_coef * 100) - 1) * 100
percent_change
```

```
## [1] 4.068032
```

With the square foot coefficient of .0003987465, we can use the above equation to find the percent change of the price when square footage increases by 100 square feet. Price is expected to change by 4.068032% based on Model 2.

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.*

```
model3 <- linear_reg() |>
  set_engine("lm") |>
  fit(price ~ log(sqft), data = houses)

model3_coeff <- tidy(model3)
print(model3_coeff)
```

```
## # 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 log(sqft)     528647.     4651.     114.        0
```

```
delta_logsqft <- log(1.10)

logsqft_coef <- model3_coeff$estimate[model3_coeff$term == "log(sqft)"]

price_change <- logsqft_coef * delta_logsqft
price_change
```

```
## [1] 50385.48
```

The value $50,358.48 represents the expected change in price when the square footage increases by 10%.

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:

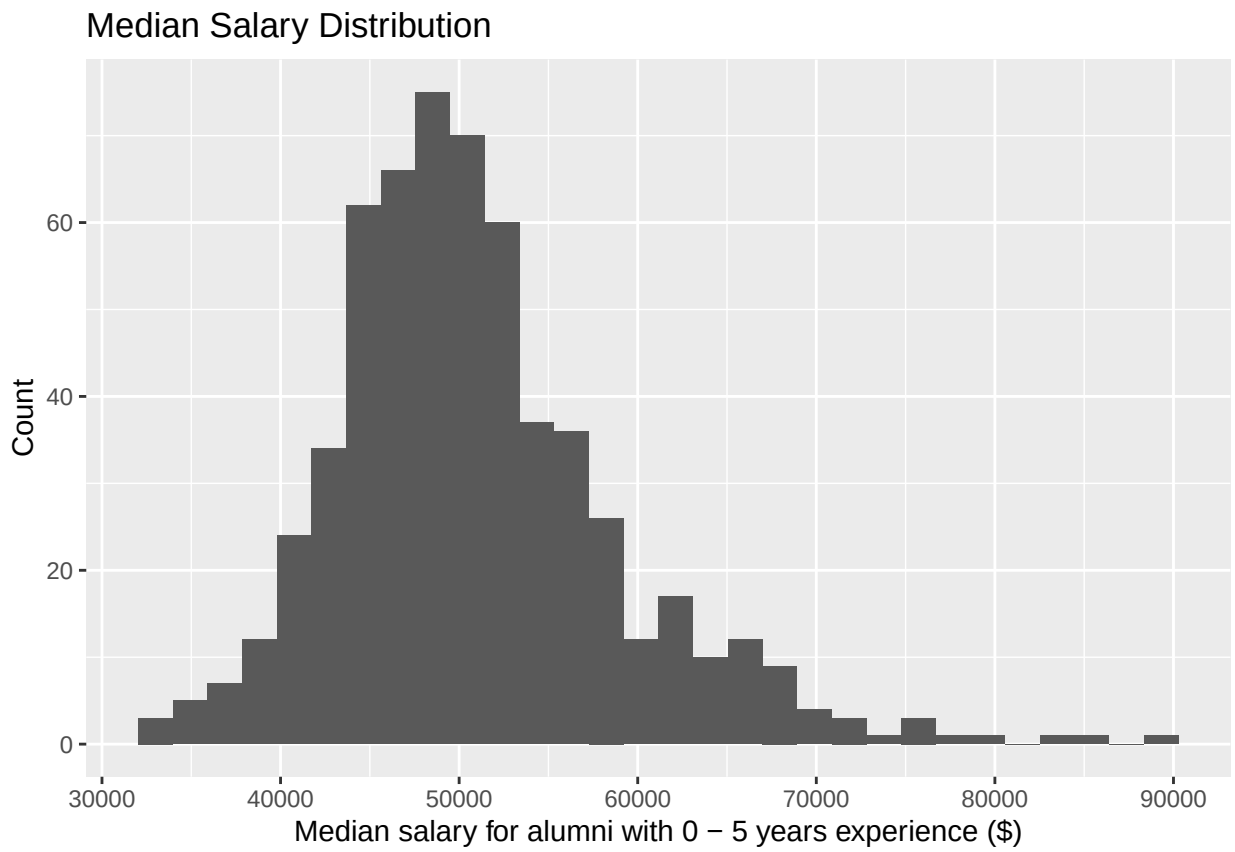
variable	class	description
name	character	Name of school
state_name	character	state name

variable	class	description
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.

```
collegepay <- read_csv("~/STA310/college-data.csv")

ggplot(data = collegepay, aes(x = early_career_pay)) +
   geom_histogram() +
     labs(x = "Median salary for alumni with 0 - 5 years experience ($)",
        y = "Count",
        title = "Median Salary Distribution")
```




```
collegepay |>
   skim(early_career_pay) |>
   select(-skim_type, -skim_variable, -complete_rate,
```

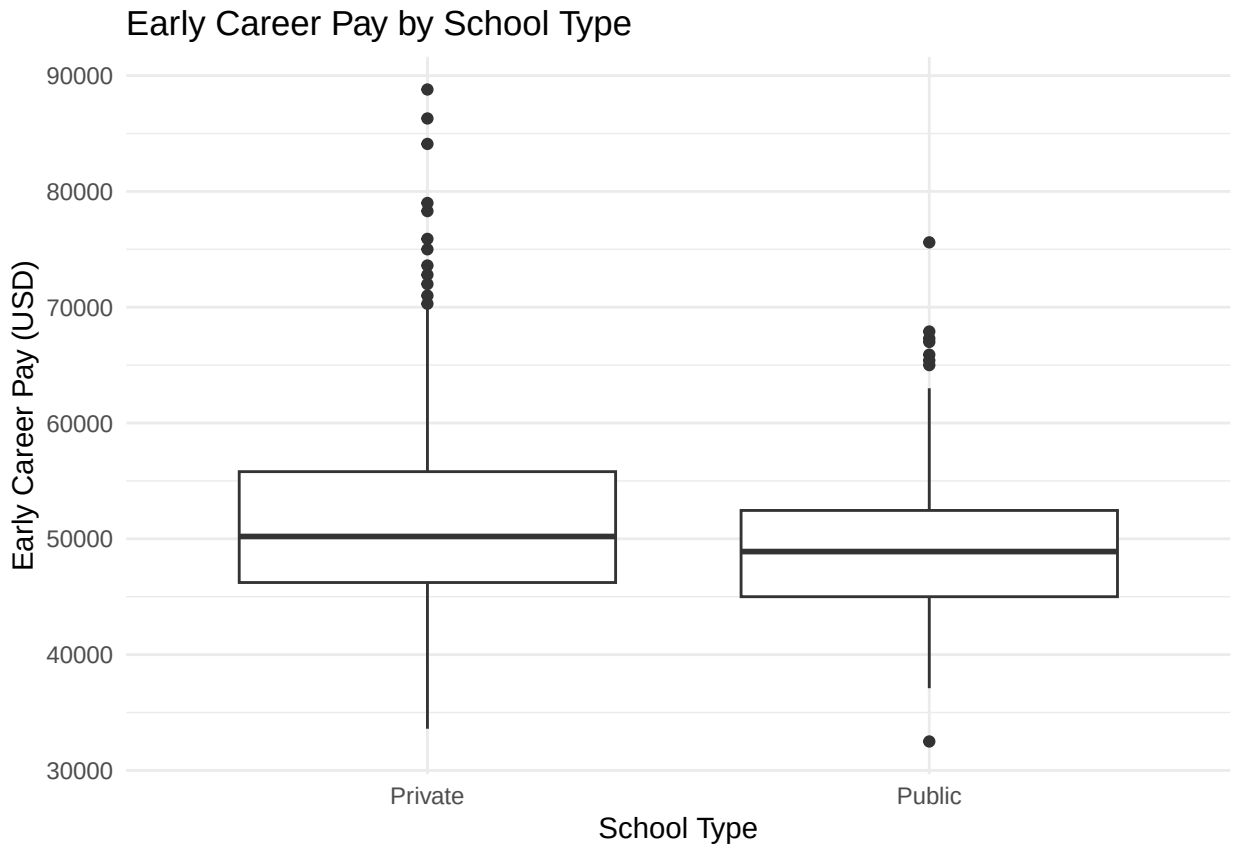
```
            - numeric.hist) |>
    print(width = Inf)
```

```
## # A tibble: 1 x 8
##   n_missing numeric.mean numeric.sd numeric.p0 numeric.p25 numeric.p50
##       <int>        <dbl>      <dbl>      <dbl>       <dbl>       <dbl>
## 1         0       50893.      7929.      32500       45900       49700
##   numeric.p75 numeric.p100
##         <dbl>        <dbl>
## 1       54500        88800
```

This is a right-skewed distribution with a mean income of $50,892.58.

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.

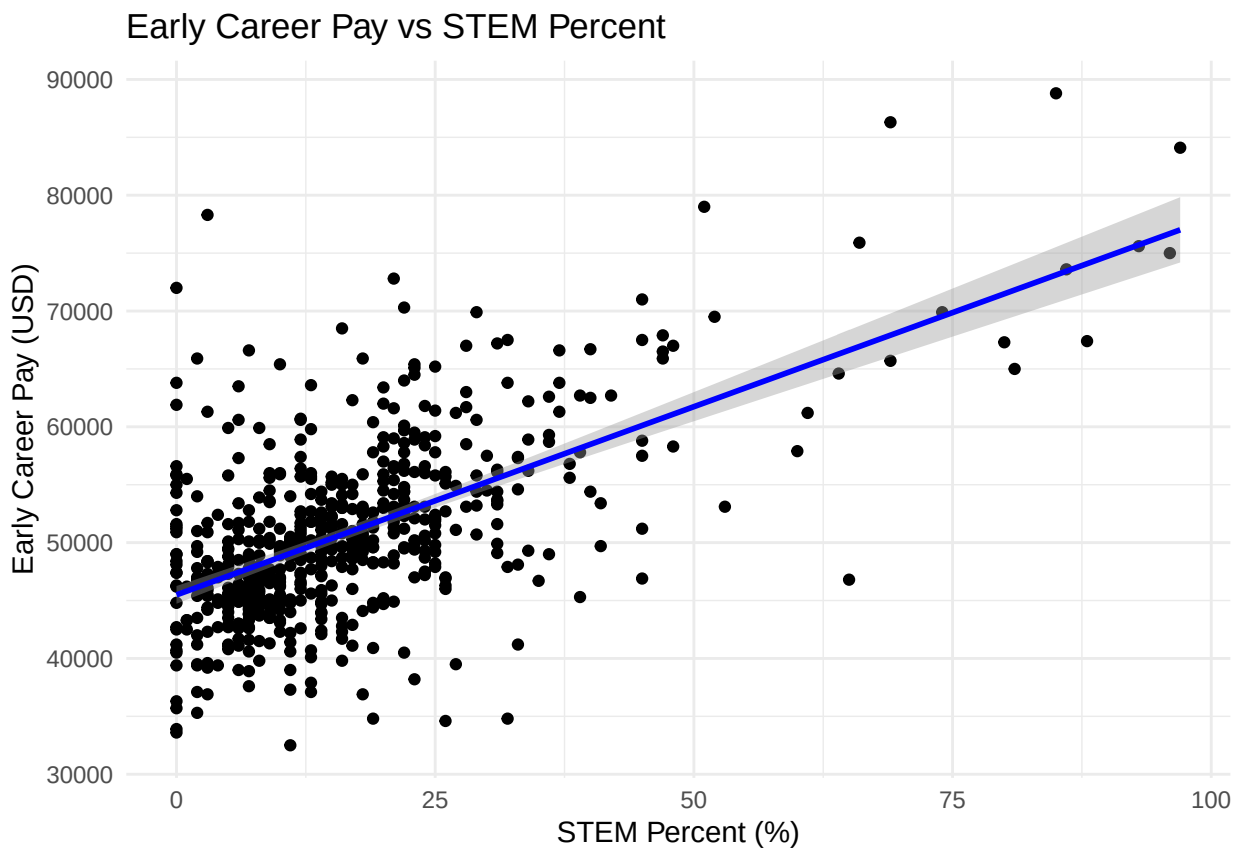
```
ggplot(data = collegepay, aes(x = type, y = early_career_pay)) +
  geom_boxplot() +
  labs(title = "Early Career Pay by School Type",
       x = "School Type",
       y = "Early Career Pay (USD)") +
  theme_minimal()
```



The median early career pay for private schools is slightly greater than public schools, while the range and variability is also greater for private schools.

```r
ggplot(data = collegepay, aes(x = stem_percent, y = early_career_pay)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Early Career Pay vs STEM Percent",
       x = "STEM Percent (%)",
       y = "Early Career Pay (USD)") +
  theme_minimal()
```

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



As the percentage of degrees rewarded in the STEM fields increases, the early career pay increases, on average.

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
library(broom)

model <- lm(early_career_pay ~ out_of_state_total + type + stem_percent +
              type:stem_percent, data = collegepay)

model_results <- tidy(model, conf.int = TRUE, conf.level = 0.95)

model_results$estimate <- round(model_results$estimate, 3)
model_results$std.error <- round(model_results$std.error, 3)
model_results$statistic <- round(model_results$statistic, 3)
model_results$p.value <- round(model_results$p.value, 3)
model_results$conf.low <- round(model_results$conf.low, 3)
model_results$conf.high <- round(model_results$conf.high, 3)

print(model_results)
```

```
## # A tibble: 5 x 7
##   term                    estimate std.error statistic p.value conf.low conf.high
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)             3.62e+4   850.        42.6    0      3.45e+4 37888.
## 2 out_of_state_total      2.53e-1     0.018     13.7    0      2.17e-1     0.289
## 3 typePublic              1.19e+3   769.         1.54   0.124 -3.25e+2  2695.
## 4 stem_percent            2.14e+2    19.3       11.1    0      1.76e+2   252.
## 5 typePublic:stem_perce~  4.95e+1    33.9        1.46   0.144 -1.70e+1   116.
```

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

The degrees of freedom can be solved by subtracting the number of parameters in the model by the number of observations in the data. In this case, there are 5 parameters and 593 observations, so there are 588 degrees of freedom.

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

```r
model_results <- tidy(model, conf.int = TRUE, conf.level = 0.95)

type_public_coeff <- model_results[model_results$term == "typePublic", ]

type_public_coeff[c("conf.low", "conf.high")]
```

```
## # A tibble: 1 x 2
##   conf.low conf.high
##      <dbl>     <dbl>
## 1    -325.     2695.
```

[-324.8133, 2694.853] is the 95% confidence interval for the amount in which the intercept for public institutions differs from private institutions.

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).*

It appears that the variables that indicate a significantly significant characteristic of the colleges and universities are the total cost for in state residents and the percentage of degrees in STEM. These two variables have a 95% confidence interval that do not include 0, while the variable indicating the type of school has a 95% confidence interval that includes 0. As the percentage of STEM degrees and the price of the college increase, the expected early career pay for graduates increases. On the other hand, the difference between a Public and Private school does not provide a statistically significant difference to early career pay.

Grading

Total	**50**
Ex 1	8
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.