STA310 HW1

Olivia Fu

2025-01-20

```
library(tidyverse)
library(tidymodels)
library(knitr)
library(ggplot2)
```

Exercise 1

(a)

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

The predictor variable is temperature at the recorded time.

(b)

$$y_i = \mu_i + \epsilon_i = x_i^T\beta + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

More specifically:

$$\text{number of cricket chirps per minute}_i = \beta_0 + \beta_1 \text{ temperature}_i + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

y_i is the i-th observation of the number of cricket chirps per minute

β_o is the intercept, which means the expected number of chirps per minute when the temperature is 0

β_1 is the slope, which represents the change in chirps per minute when the temperature increase by 1 unit

ϵ_i is the error term that counts the random part of the model, and they are independent and identically distributed $N \sim (0, \sigma^2)$

(c)

Linearity: The relationship between mean of the number of cricket chirps per minute (response Y) and temperature at the recorded time (predictor X) is linear.

Independence: Each observation of the cricket chirps and temperature pair is independent of the others. There is no connection between how far any two data points lie from the regression line.

Normality: The number of cricket chirps per minute (response Y) follows a normal distribution at each level/value of temperature (predictor X).

Equal variance: Variance of the number of cricket chirps per minute (response Y) is constant across all values of temperature.

Exercise 2

(a)

The response variable is postnatal depression, specifically patients' depression scores.

The predictor variable is whether or not an estrogen patch is used.

(b)

Violation of Independence: Based on the scenario, the assumption of independence is violated because patients' depression scores were recorded on 6 different visits. The depression scores of the same group of patients were recorded for multiple times. Therefore, the observations, particularly those from the same patients, are correlated and not independent of each other.

Normality Assumption: We need additional information about the data to determine whether the normality assumption is satisfied. If the distribution of depression scores is approximately normal for both groups, the assumption holds. However, if all patients have similar depression scores or if the distribution is skewed (e.g., probably due to a smaller chance of observing severely depressed patients), the normality assumption would be violated.

Exercise 3

(a)

In this model, we include year as an additional predictor variable alongside track conditions. As observed in the exploratory data analysis before, winning speed varies over time, indicating that year has an impact on winning speed. This model allows us to estimate the difference in winning speeds between fast and non-fast track conditions for a given year. By doing so, it separates the effect of track conditions from the trends over time (year). When interpreting β_2, it represents the effect of track conditions on winning speed while controlling the time factor. Therefore, it's important to state "holding year constant".

(b)

The equation provides the predicted values of Y_i based on the fitted regression model. This is a regression equation that estimates the function based on the sample data. The error term accounts for random variation not explained by the deterministic component of the model. However, since the regression equation focuses on predicted values, it does not include the potential deviations of actual observations from the predicted values.

Exercise 4

(a)

```r
house <- read.csv("~/Desktop/STA310/sta310-spring25/HW1/kingCountyHouses.csv")
```

```r
lm(price ~ sqft, data = house) |>
  tidy() |>
  kable(digits = 4)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-43580.7431	4402.6897	-9.8987	0
sqft	280.6236	1.9364	144.9204	0

The slope coefficient of model 1 is 280.6236. In this context, for every one unit increase in the interior square footage, the selling price of the house is expected to increase by 280.6236 dollars, on average. Therefore, when the interior square footage (sqft) increases by 100, we expect the selling price of the house to increase by 28062.36 dollars, on average.

(b)

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

lm(logprice ~ sqft, data = house) |>
  tidy() |>
  kable(digits = 4)
```

term	estimate	std.error	statistic	p.value
(Intercept)	12.2185	0.0064	1916.8830	0
sqft	0.0004	0.0000	142.2326	0

The slope coefficient of model 2 is 0.0004. In this context, for every one unit increase in the interior square footage, the log price of the house is expected to increase by 0.0004, on average. Therefore, when the interior square footage (sqft) increases by 100, the log of the house price is expected to increase by 0.04, on average.

(c)

$log(\text{new price}) - log(\text{old price}) = log(\frac{\text{new price}}{\text{old price}}) = 0.04$

$\frac{\text{new price}}{\text{old price}} = e^{0.04} = 1.0408$

Based on model 2, when the interior square footage (sqft) increases by 100, the house price is expected to multiply by a factor of 1.0408 (exp(0.04)), on average.

(d)

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

lm(price ~ logsqft, data = house) |>
  tidy() |>
  kable(digits = 4)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-3451377.1	35169.35	-98.1359	0
logsqft	528647.5	4650.63	113.6722	0

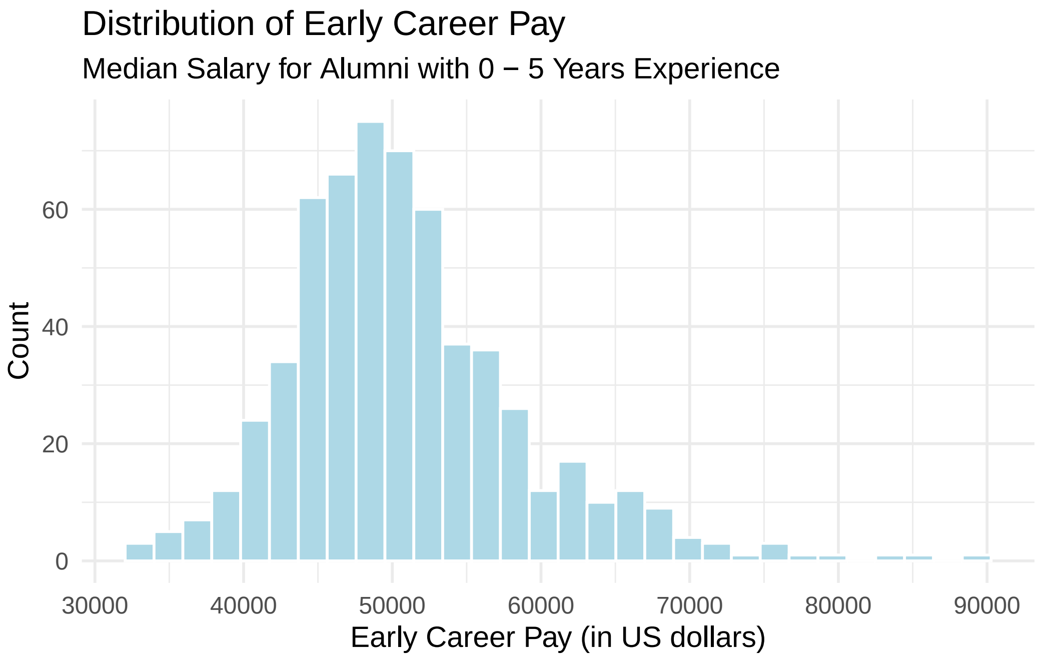
The slope coefficient of model 3 is 528647.5. In this context, a 10% increase in interior square footage (sqft) corresponds to the logsqft to increase by 0.09531 (log(1.1)). Therefore, when sqft increases by 10%, the house price is expected to increase by $528647.5 \times 0.09531 = 50385.39$ dollars, on average.

Exercise 5

(a)

```
college <- read.csv("~/Desktop/STA310/sta310-spring25/HW1/college-data.csv")
```

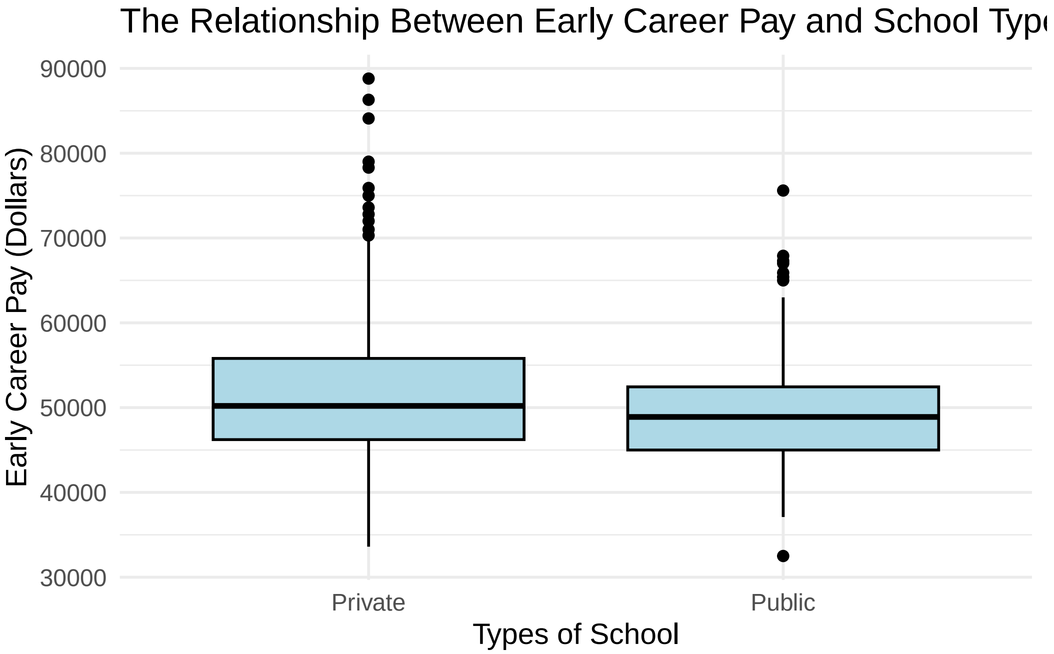
```
ggplot(data = college, aes(x = early_career_pay)) +
  geom_histogram(fill = "lightblue", color = "white") +
  labs(x = "Early Career Pay (in US dollars)",
       y = "Count",
       title = "Distribution of Early Career Pay",
       subtitle = "Median Salary for Alumni with 0 - 5 Years Experience") +
  theme_minimal()
```



The early career pay of alumni follows a slightly **right-skewed** distribution. The distribution is centered around $48,000, with frequencies gradually decreasing on both sides. Most data points (alumni's early career pay) fall in the range of $40,000 and $60,000. There are a few data points with very high early career pay, resulting in a long tail on the right side of the distribution.

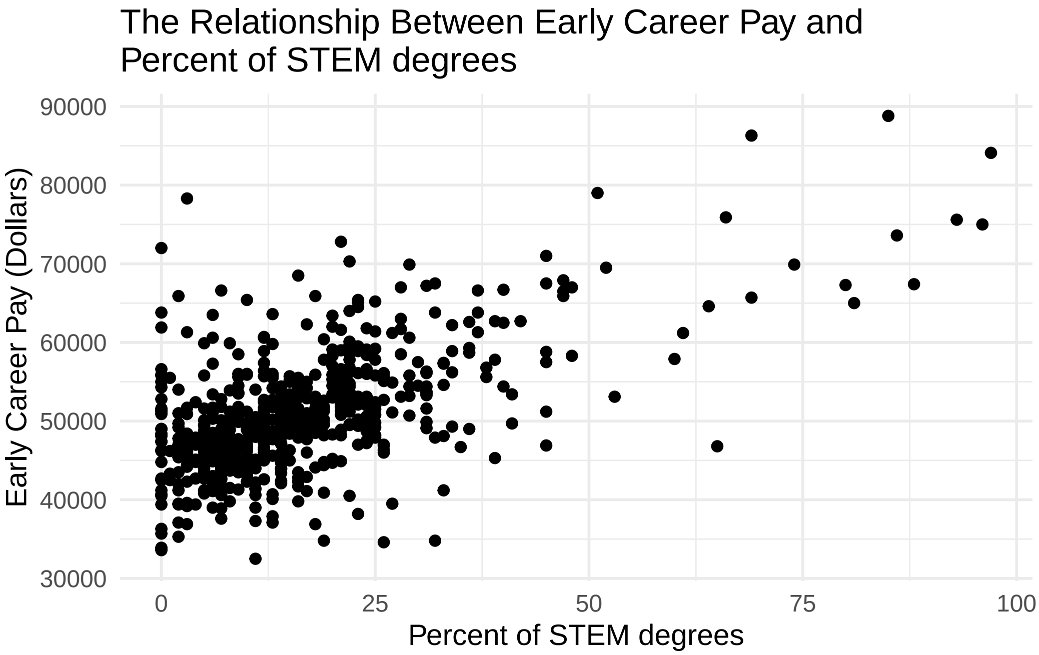
(b)

```
ggplot(data = college, aes(x = type, y = early_career_pay)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(x = "Types of School",
       y = "Early Career Pay (Dollars)",
       title = "The Relationship Between Early Career Pay and School Type") +
  theme_minimal()
```



As observed from the graph, the median of early career pay for private school alumni is slightly higher than that of students graduated from public schools. Furthermore, a greater number of private school alumni have an early career pay exceeding $70,000 compared to public school graduates.

```
ggplot(data = college, aes(x = stem_percent, y = early_career_pay)) +
  geom_point() +
  labs(x = "Percent of STEM degrees",
       y = "Early Career Pay (Dollars)",
       title = "The Relationship Between Early Career Pay and
Percent of STEM degrees") +
  theme_minimal()
```



As observed from the graph, there is a positive relationship between the percentage of STEM degrees awarded and alumni's early career pay, which means that alumni from schools awarding a higher percentage of STEM degrees tend to have higher early career pay. Most data points are concentrated between 0% and 25% STEM degrees, while those above are more scattered.

(c)

```
lm(early_career_pay ~ out_of_state_total + type + stem_percent +
     type * stem_percent,
   data = college) |>
  tidy(conf.int = TRUE, conf.level = 0.95) |>
  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
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)

There are $n - p - 1 = 593 - 4 - 1 = 588$ degrees of freedom in the estimate of the regression standard error.

(e)

The 95% confidence interval for the amount in which the intercept for public institutions differs from private institutions is [-324.813, 2694.853].

Exercise 6

Our analysis shows that certain institutional characteristics influence alumni’s early career pay; specifically, the total cost for out-of-state residents and the percentage of STEM degrees awarded are associated with higher early career salaries. For every $1 increase in the total cost for out-of-state residents, alumni’ early career pay is expected to increase by $0.253, suggesting that institutions with higher tuition tend to have graduates with slightly higher salaries during their first five years of work. Additionally, a 1% increase in the percentage of STEM degrees awarded leads to the alumni’s early career pay to increase by $214.31, indicating that STEM education can positively impact graduates’ early career pay. On the other hand, whether an institution is public or private does not significantly affect early career pay, as both the school type and its interaction with STEM degree percentage have slopes with confidence intervals that encompass zero. These findings highlight that institutional cost and STEM education have a stronger impact on alumni’s early career pay, while the type of school (public vs. private) does not play a significant role.