Stat 3301: Homework 3

FirstName LastName (name.n)

Due date set on Carmen

Setup:

```r
library(alr4)
```

```
## Warning in check_dep_version(): ABI version mismatch:
## lme4 was built with Matrix ABI version 1
## Current Matrix ABI version is 0
## Please re-install lme4 from source or restore original 'Matrix' package
```

```r
library(tidyverse)
library(readr)
```

Instructions

- Replace "FirstName LastName (name.n)" above with your information.
- Provide your solutions below in the spaces marked "Solution:".
- Include any R code that you use to answer the questions; if a numeric answer is required, show how you calculated it in R.
- Knit this document to **HTML** and upload both the **HTML** file and your completed Rmd file to Carmen
- Make sure your solutions are clean and easy-to-read by
  - formatting all plots to be appropriately sized, with appropriate axis labels.
  - only including R code that is necessary to answer the questions below.
  - only including R output that is necessary to answer the questions below (avoiding lengthy output).
  - providing short written answers explaining your work, and writing in complete sentences.
- Data files mentioned below are from the `alr4` package unless specified otherwise.

Concepts & Application In this assignment, you will

- write down the components and assumptions of a simple linear regression model
- identify the mean and variance functions of a SLR model
- estimate the parameters of a SLR model using summary statistics also using the `lm` function, and interpret the results
- plot an estimated simple linear regression line
- compare two fitted regression models
- use a fitted model to describe relationships between variables
- assess whether the SLR model is an appropriate model

Question 1 When asked to state the simple linear regression model, a student wrote is as follows: $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_1 + e_i$. Do you agree? Please comment on this.

Solution to Question 1

I do not agree. Instead of X_1, it should be X_i.

Question 2 This question relates to the **SunTec** data set which can be found on Carmen.

SunTec is a manufacturer of industrial auto parts products. The researchers at SunTec want to investigate the study of relationship between length of employment (in months) of their employees and number of auto parts sold by the employees. `Data source: Statistics for Business and Economics, 13th Edition, by Anderson, Sweeney,Williams, Camm and Cochran.`

2.1. Using `R`, Draw a scatter plot of $y = ItemSold$ versus $x = Months$.

2.2. On this plot, draw the ols fitted line (estimated regression line).

2.3. Give two reasons why fitting simple linear regression to this problem is not likely to be appropriate.

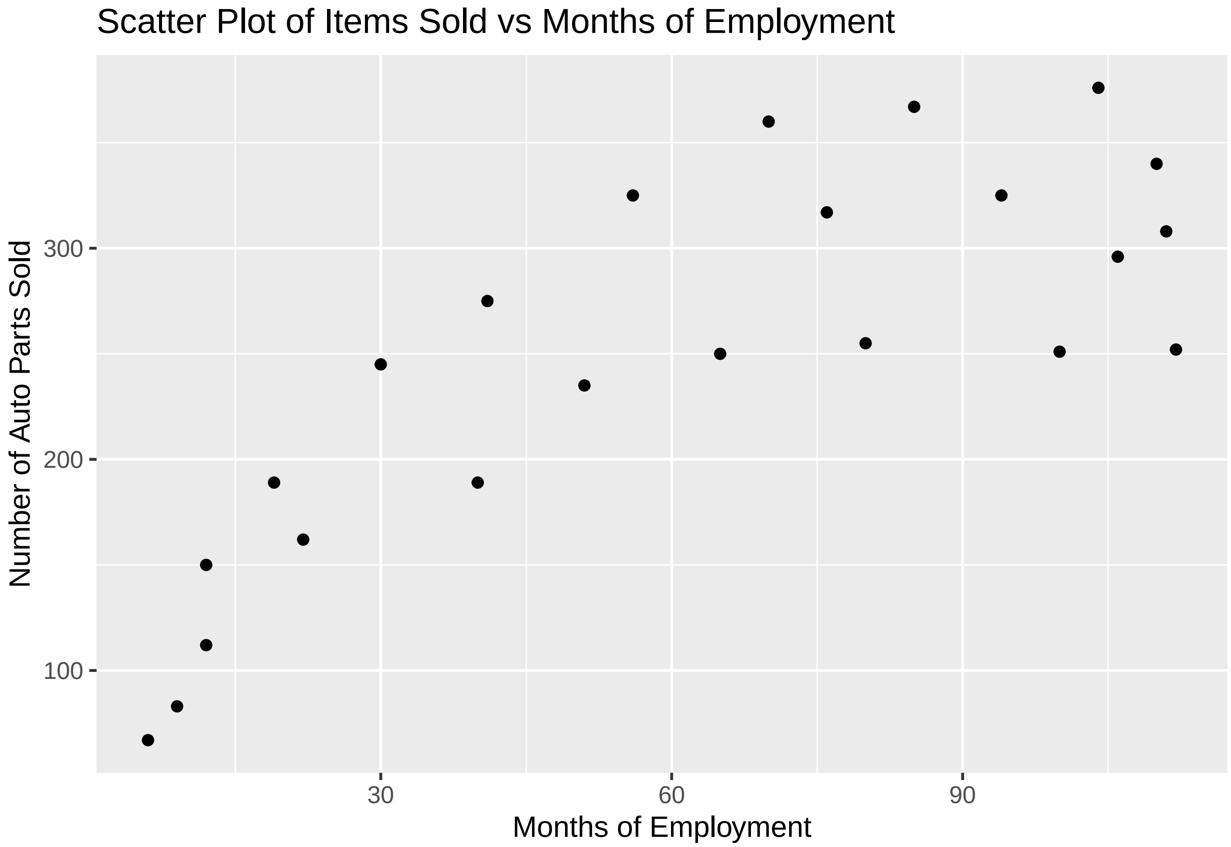
Solution to Question 2 Your answers go here.

Part 2.1:

```r
suntec_data <- read_csv('SunTec.csv')
```

```
## Rows: 23 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Months, ItemSold
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
ggplot(suntec_data, aes(x = Months, y = ItemSold)) + geom_point() +
  labs(title = "Scatter Plot of Items Sold vs Months of Employment",
       x = "Months of Employment",
       y = "Number of Auto Parts Sold")
```



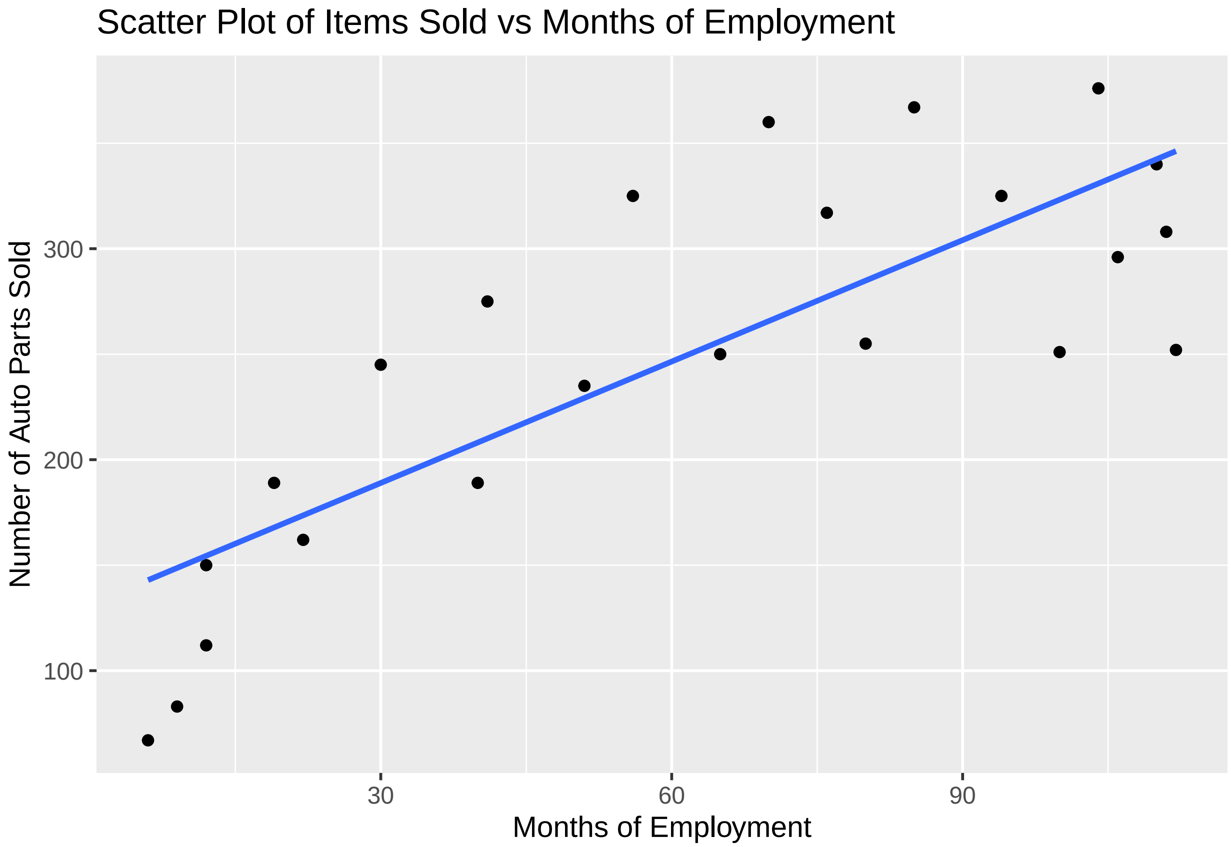
Part 2.2:

```r
suntec_data <- read_csv('SunTec.csv')
```

```
## Rows: 23 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): Months, ItemSold
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
ggplot(suntec_data, aes(x = Months, y = ItemSold)) + geom_point() +
  labs(title = "Scatter Plot of Items Sold vs Months of Employment",
       x = "Months of Employment",
       y = "Number of Auto Parts Sold") + geom_smooth(method = "lm", se = F)
```

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



Part 2.3: Fitting simple linear regression to this problem is not likely because the data is not linear, and there is not constant variance with each amount of months of employment.

Question 3 In a simple linear regression (SLR), it was derived in the class that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SXY}{SXX}$$

Please show that

$$\hat{\beta}_1 = r_{xy}\frac{SD_y}{SD_x}$$

Please refer Table 2.1 from the Textbook for the definition of the symbols used in this problem.

Solution to Question 3 $r_{xy} = \frac{s_{xy}}{(SD_x SD_y)}$ $\frac{s_{xy}}{SD_x SD_y} \cdot \frac{SD_y}{SD_x} = \frac{S_{xy}}{SD_x^2} = \frac{s_{xy}}{SXX/(n-1)} = \frac{s_{xy} \cdot (n-1)}{SXX} = \frac{SXY}{n-1} \cdot \frac{n-1}{SXX} = \frac{SXY}{SXX} = \hat{\beta}_1$

Question 4 This question relates to the **PROSalary** data set which can be found on Carmen.

This data frame contains the following columns:

- salary - dollars per year
- exper – Years of experience in the firm
- degree - individual has a relevant graduate degree with levels "N" or "Y"
- degreetype - Factor with levels "BS", "PHD" or "MS"
- score - score on the programmer aptitude test

A software firm collected data for a sample of 23 computer programmers. A suggestion was made that regression analysis could be used to determine study of relationship between salary and the years of experience.

Here we will focus on programmers who have **no relevant degree** (labeled `N` in the data set).

1. Make a plot of programmer salary vs. exper for all programmers who have **no relevant degree**. Use the plot to summarize the relationship between salary and exper for programmers who have no relevant degree in the firm.
2. We will use the normal simple linear regression model to relate salary to exper for these data (salary is the response variable). Write down the general form for the model, starting out with:

   $$salary_i = \beta_0 + \cdots, \quad i = 1, \ldots, n.$$

   Continue to fill out the rest of the right hand side of the equation. Your model should be expressed in terms of unknown parameters (e.g., β_1) and not specific estimated values (e.g.$\widehat{\beta}$) or numbers (e.g., 0.23). Make sure you include an error term, e_i, in the model and **be sure to specify all assumptions about its distribution**.
3. Write down the conditional mean function $E(Y \mid X)$ and the conditional variance function $\mathrm{Var}(Y \mid X)$ for salary given exper as a function of the unknown parameters. (You will estimate the parameters next.)
4. Calculate the summary statistics ($\bar{x}$, SXX, etc.) required to compute the ordinary least squares estimates of the parameters β_0 and β_1 in the mean function, and use these statistics to calculate the estimated values $\widehat{\beta}_0$ and $\widehat{\beta}_1$.
5. Use the `lm` function in `R` to compute the least squares estimates and display the results. Also, compare with your results above (they should be the same).
6. Calculate an unbiased estimate of the conditional variance of salary given exper.

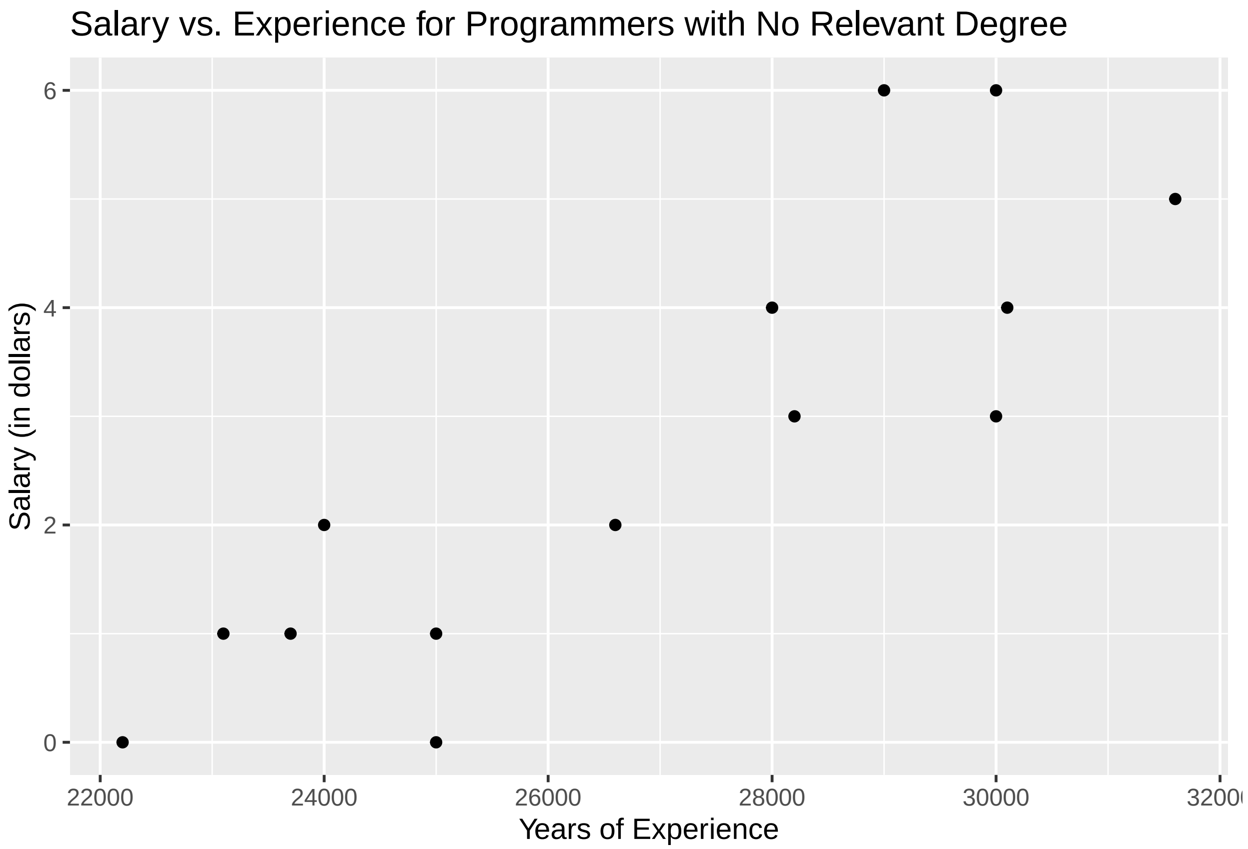
Solution to Question 4 Your answers go here.

Part 4.1:

```
prosalary_data <- read_csv('PROSalary.csv')
```

```
## Rows: 23 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): degree, degreetype
## dbl (3): exper, score, salary
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
no_degree_data <- prosalary_data %>% filter(degree == "N")
ggplot(no_degree_data, aes(x = salary, y = exper)) + geom_point() + labs(title = "Salary vs. Experience
       x = "Years of Experience",
       y = "Salary (in dollars)")
```



This plot shows a positive relationship between years of experience and salary. As the years of experience increase, so does the salary.

Part 4.2:

$$salary_i = \beta_0 + \beta_n \cdot exper_i + e_i, i = 1, 2, \ldots, n.$$

e_i is independent for each observation and e_i follows a normal distribution.

Part 4.3:

$$E(salary_i \mid exper_i) = \beta_0 + \beta_n \cdot exper_i$$

$$Var(salary_i \mid exper_i) = \sigma^s$$

Part 4.4:

```
exper <- no_degree_data$exper
salary <- no_degree_data$salary
```

```
x_bar <- mean(exper)
y_bar <- mean(salary)

SXX <- sum((exper - x_bar)^2)
SXY <- sum((exper - x_bar) * (salary - y_bar))

beta1_hat <- SXY / SXX
beta0_hat <- y_bar - beta1_hat * x_bar

beta0_hat
```

```
## [1] 23425.78
```

```
beta1_hat
```

```
## [1] 1277.344
```

Part 4.5:

```
lm_model <- lm(salary ~ exper, data = no_degree_data)

coef(lm_model)
```

```
## (Intercept)       exper
##   23425.781    1277.344
```

Part 4.6:

```
lm_model <- lm(salary ~ exper, data = no_degree_data)

summary(lm_model)$sigma^2
```

```
## [1] 2653665
```