STA238 Final Project

By Henry Jia Yang Lu / William Sung Chi Wu

Introduction

Attrition is a headache for all businesses, regardless of the industry, geography and scale of the company. Employee attrition can cause great costs for a company, such as the cost of hiring new staff, training new employees and business disruptions. Therefore, there is great interest in understanding the drivers of attrition, and minimizing it. For this project, we are analyzing a dataset originated from an IBM employee survey including employee's general information, career status, and if there is attrition or not. Using this dataset, we came up with two research questions that we are going to study with various methods that we've accumulated throughout the course.

Research Question #1: Do employees that attrit have lower average monthly income and job satisfaction compared to those that do not?

Research Question #2: Is there any association between a male IBM sales employee's total years worked and their monthly income?

Our dataset contains the data of 1470 IBM employees and 35 variables. The main targeted variables for this research is "Attrition" which has possible outcomes Yes and No. "MonthlyIncome ", which represents an employee's monthly wage in dollars. And "TotalWorkingYears', which represents the total number of years in an employee's career(including working years outside of IBM). The 32 other variables are independent except for EmployeeNumber, which represents each employee's unique identification number. The data does not include frequency of data collection or date range. For data cleaning, we removed all the values that are "not applicable" from the three variables that we mentioned previously. To set up for our first research question, we've created two datasets based on the cleaned data to separate employees who plan on attriting and who remain in IBM; the datasets are named "stay" and "leave." For our second research question, we narrowed the dataset to only include male employees who work in the sales department of IBM. The purpose of this is to exclude the factors that may potentially interfere with our result, therefore we are specifically studying the association between monthly wage and total work years of the male employees within the sales

department (Appendix 1.1). We also performed a few exploratory data analysis before conducting our research

Exploratory Data Analysis

Numerical Summaries of Monthly Income for Attrition Employees(Appendix 1.2)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1051	3211	5204	6833	8834	19999

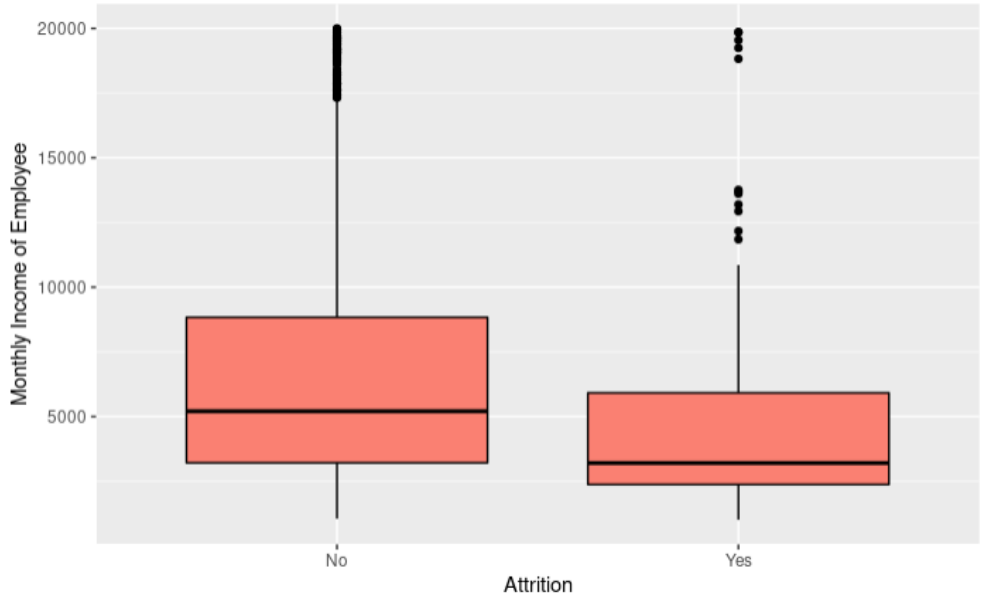
1. Numerical Summaries of Monthly Income for Non-Attrition Employees(Appendix 1.2)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1009	2373	3202	4787	5916	19859

2. Side By Side Boxplot of Monthly Income for both employee groups(Appendix 1.3)

   (X-Axis: Attrition = Yes and Attrition = No)

   (Y-Axis: Monthly Income in US dollars)



To answer our first research question: Do employees that attrit have lower average monthly income and job satisfaction compared to those that do not, we will use the Two Group Hypothesis Testing and the Confidence Intervals for Two Population Means.

Research Question 1:Two Group Hypothesis Testing(Appendix 2.1)

We will use Two Group Hypothesis Testing to verify if there is a difference between the average income of workers who are leaving (Attrition = Yes) and that of workers who are staying (Attrition = No). The null hypothesis is set to be that the average monthly income of employees who are staying and that of the employees who are leaving are equal. Our alternative hypothesis will be that the average monthly income between the two types of employees are different.

Let:

$\mu_1 = \ average\ monthly\ income\ for\ employees\ who\ are\ leaving\ the\ IBM$

$\mu_2 \ = \ average\ monthly\ income\ for\ employees\ who\ are\ staying\ in\ IBM$

Then our null hypothesis and alternative hypothesis would be:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 < \mu_2$

For this method, we are using the two dataset “leave” and ”stay” that contains employees who are leaving (Attrition = Yes) and another that contains employees who are staying (Attrition = No). We then compute the income mean for attrition employees ($\overline{x_1}$) and for non-attrition employees ($\overline{x_2}$) and also the standard deviation for attrition employees (s_1) and for non-attrition employees (s_2). The sample size of attrition employees and non-attrition employees are represented by n_1 and n_2, respectively. After computing these parameters, the test statistic is computed by the following equation:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

The degree of freedom is calculated using the Welch’s Degree of Freedom:

$$df = \frac{[{s_1}^2/n_1 + {s_2}^2/n_2]^2}{\frac{[{s_1}^2/n_1]^2}{n_1 - 1} + \frac{[{s_2}^2/n_2]^2}{n_2 - 1}}$$

Finally, the p-value is computed in R with parameters t-value of -7.4826 and degree of freedom 412. The result for our p-value is $2.223421e^{-13}$.

Since the p-value is less than 0.01, which tells us that under the original assumptions, an observation of the difference between the monthly income of the attrition and non-attrition employees has a low probability of occurrence. There is strong evidence against our null hypothesis, and the alternative hypothesis is accepted as the result is highly statistically significant.

Research Question 1:Two Population Mean Bootstrapped Confidence Interval(Appendix 2.2)
To find out the approximate difference between the two group's monthly income, we are going to estimate the difference in average monthly income for attrition employees and non-attrition employees with 95% confidence using bootstrapped methods, since the distribution type of monthly income and true variance is unknown.

We first draw two bootstrap samples with replacement, one from the attrition group and another from the non-attrition group. We then compute the difference between the means for the bootstrap samples of the two groups. To obtain the bootstrapped sampling distribution, the process is repeated 5000 times. Finally, the percentile method is used to construct the 95%-level confidence intervals, where the lower bound and upper bound are taken to be the 0.025 and 0.975 sample quantiles of the bootstrap sampling distribution, respectively.

Our findings suggest that the 95% confidence interval of mean difference in monthly wage between attrition employees and non-attrition employees is between $1520.94 and $2568.75.

Combining the findings from the two of our methods, we can conclude that the monthly income of attrition employees are on average $1520.94 to $2568.75 lower than non-attrition employees.

We will be using Simple Linear Regression (SLR) to answer our second research question: Is there any association between an IBM sales employee's total years worked and their monthly income. The results of SLR will determine whether the two variables are associated and the strength of their association, if it exists.

Research Question 2: Residual Analysis
Before performing the regression analysis, We will verify the following model assumptions first:

1. **<u>Random noise/residual follows normal distribution(Appendix 3.1):</u>**
   We plotted a histogram and normal QQ plot to verify the normality of the residuals. The histogram shows a normal distribution because it is unimodal centered at 0 and has a symmetric spread. The normal QQ plot also suggests a normal distribution

since the majority of the points are on the line with slight deviation on both ends. Therefore, the normality assumption holds.

2. **Average noise is zero/Residual have a mean of 0(Appendix 3.2, 3.3):**
   We plotted the residual horizontally against our x variable(Total_work_years), and it appears to have equal amounts of positive residual and negative residuals. Furthermore, we computed the rounded mean residual which equals to 0. Therefore this assumption holds.

3. **Independence of residuals(Appendix 3.2):**
   From the horizontal residual plot, there is no obvious trend between the residual and the x variable. The random noise seems to occur independently to each other, therefore the assumption holds.

4. **Constant variance of residuals across all x variable(Appendix 3.2):**
   Since the spread of the residuals at each x variable(Total_work_years) seems to be fairly constant, this assumption holds.

Therefore, since we have verified all four assumptions, we are able to perform and use inferences we make on the estimated model that we are going to fit.

Research Question 2: Simple Linear Regression

The scatter plot shows a positive linear relation with moderate strength(Appendix 4.1). Meaning that as the total work years of an employee increase, their monthly wage also increases.
From the summary of this model(Appendix 4.2), we can determine that the fitted equation is:

$$\widehat{y_i} = 2470.11 + 413.37 * x_i$$

,where $\widehat{y_i}$ is monthly income and x_i is total work years.

The model tells us that when the employee has no work experience, their estimated monthly wage is $\beta_0 = \$2470.11$. Every additional year of work experience an employee has will increase their estimated monthly wage by $\beta_1 = \$413.37$. Therefore, experienced employees are likely to be paid more monthly.
The model also tells us that $R^2 = 0.5657$, which means that about 56.57% of total variability in an employee's monthly income can be explained by their total work years. The p-value for testing the null hypothesis H_0: $\beta_1 = 0$, and the alternative hypothesis H_A: $\beta_1 \neq 0$, is less than $2.2e^{-16}$, which is less than 0.001. This is highly statistically significant so we reject the null hypothesis. The test result concludes that there is a linear correlation between monthly income and total work years.

The 95% confidence interval for the intercept is [1892.255, 3047.965] and for the slope is [368.706, 458.034] (Appendix 4.3). The interval for the intercept is relatively wide, meaning that the estimated monthly income for employees with no work experience can be quite inconsistent. The interval for the slope is much narrower, therefore the increase in monthly wage as total work year increases by one is relatively more consistent.

Conclusion

In conclusion, this research attempted to answer the following research questions:

1. Do employees that attrit have lower average monthly income and job satisfaction compared to those that do not?
2. Is there any association between a male IBM sales employee's total years worked and their monthly income?

Two Group Hypothesis Testing and Two Population Mean Bootstrapped Confidence Interval are conducted to answer our first research question. With a p-value of $2.223421e^{-13}$, our hypothesis testing concludes that the average monthly income for employees who are leaving IBM is evidently lower than for employees who are staying. To estimate the difference between the monthly income between the two groups, we constructed a confidence interval for the difference in means between the two groups using a bootstrap approach. The percentile method is used to construct the 95%-level confidence interval, which concludes that the monthly income of attrition employees are on average $1520.94 to $2568.75 lower than non-attrition employees. Though the method is simple as it makes few assumptions about the shape of the bootstrap sampling distribution, it is important to note that the resulting interval may be to some extent biased.

Simple Linear Regression (SLR) is conducted to answer our second research question. To validate the SLR model, we conducted a residual analysis and verified that the four model assumptions hold for our data set. Therefore, the confidence interval and hypothesis test results should be valid. However, constant variance assumption is based on the assumption that at each total working years, we should see similar deviations in the residuals. However, at a total working years above 20, we are limited by the size of our data set and we may have insufficient observations at a total working years to fully capture the residual spread. With a p-value as small as $2.2e^{-16}$ and a R^2 of 52.67%, our SLR suggests that there is indeed a linear correlation with moderate strength between total work years and monthly income. Combining the scatterplot and the fitted model, we can conclude that there is a positive linear relationship of moderate strength between the monthly wage and total work years of an employee.

APPENDIX AND CITATION

Appendix 1.1

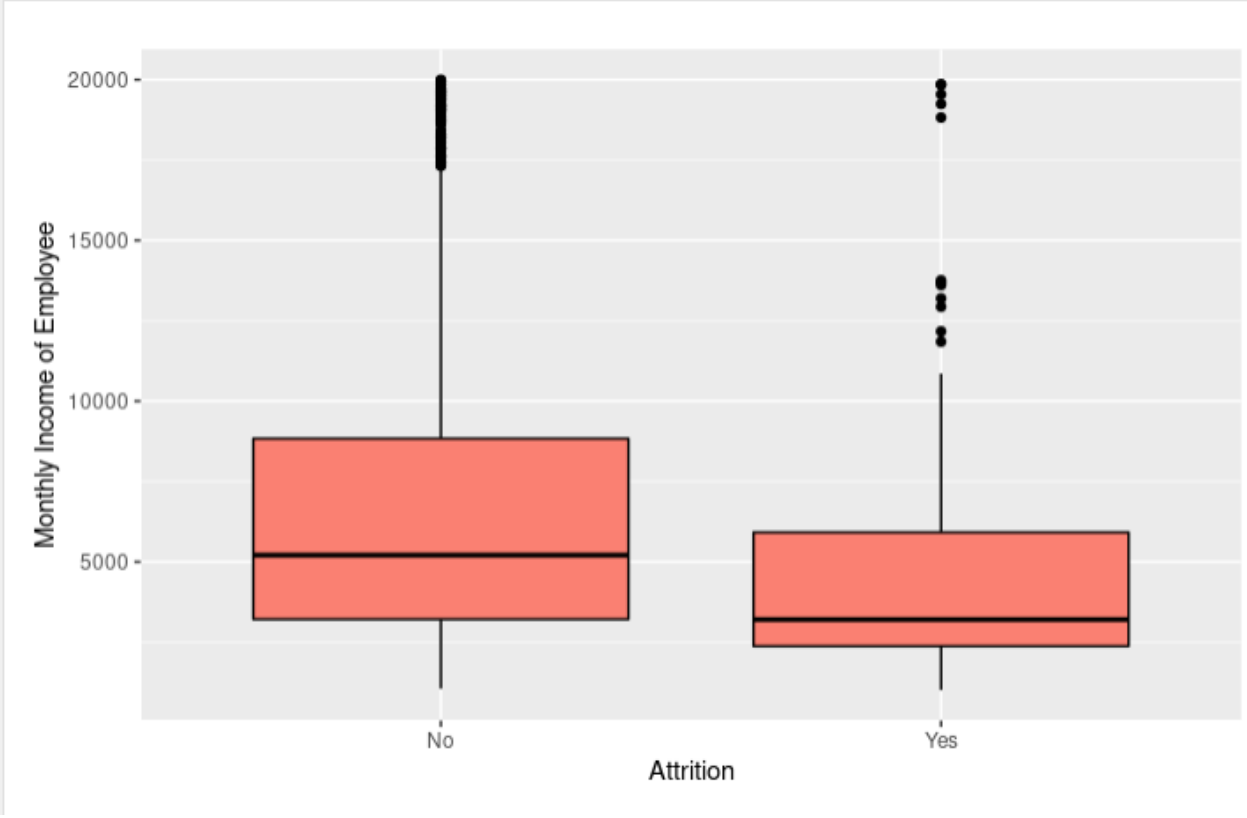
DATA CLEANING/MANIPULATION

```
```{r}
#DATA CLEANING FOR THIS PROJECT

#Read the data file
IBM_data <- read_csv("IBM HR-Employee-Attrition.csv")
#Filter out NA in the variable Attrition, MonthlyIncome and TotalWorkingYears
IBM_clean <- IBM_data%>%
  filter(!is.na(Attrition)&
         !is.na(MonthlyIncome)&
         !is.na(TotalWorkingYears))

#DATA MANIPULATION FOR RESEARCH QUESTION 1
#Create a dataset to only contain employees who are leaving IBM
leave <- IBM_clean%>%
  filter(Attrition == "Yes")

#Create a dataset to only contain employees who are staying in IBM
stay <- IBM_clean%>%
  filter(Attrition == "No")

#DATA CLEANING/MANIPULATION FOR RESEARCH QUESTION 2
#Filter the data to target only male employees in the Sales department
Male_Sales_data <- IBM_clean%>%filter(Gender == "Male" & Department == "Sales")

#Store the MonthlyIncome and TotalWorkingYears of male_sales_data into a tibble
SLR_data <- tibble(monthly_income = Male_Sales_data$MonthlyIncome,
                   total_work_year = Male_Sales_data$TotalWorkingYears)

```
```

Appendix 1.2

INITIAL EDA

```
```{r}
att_data <- IBM_clean%>%filter(Attrition == "No")
summary(att_data$MonthlyIncome)
```
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1051    3211    5204    6833    8834   19999
```

```
```{r}
no_att_data <- IBM_clean%>%filter(Attrition == "Yes")
summary(no_att_data$MonthlyIncome)
```
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1009    2373    3202    4787    5916   19859
```

Appendix 1.3

```r
ggplot(data=IBM_clean, aes(x = Attrition, y = MonthlyIncome)) +
  geom_boxplot(fill = "salmon", color = "black") +
  labs(x = "Attrition", y = "Monthly Income of Employee")
```



Appendix 2.1

```r
#Let x_bar_1 be income mean of "attrition" employees
A_income_mean <- mean(leave$MonthlyIncome)
#Let x_bar_2 be income mean of "not attrition" employees
NA_income_mean <- mean(stay$MonthlyIncome)

#Let S_1 be the standard deviation of "attrition" employees
A_sd <- sd(leave$MonthlyIncome)
#Let S_2 be the standard deviation of "not attrition" employees
NA_sd <- sd(stay$MonthlyIncome)

#Let n_1 be the number of "attrition" employees
A_size <- nrow(leave)
#Let n_2 be the number of "non attrition" employees
NA_size <- nrow(stay)

#Compute the test statistic
test_stats <- (A_income_mean - NA_income_mean)/
  sqrt(A_sd^2/A_size + NA_sd^2/NA_size)

#Compute the degree of freedom
df <- floor ((A_sd^2/A_size + NA_sd^2/NA_size)^2/
               (((A_sd^2)/A_size)^2/(A_size-1)+((NA_sd^2)/NA_size)^2/(NA_size-1)))

#Calculate the P-value(H_a: mu_1 < mu_2 )
p_value <- pt(test_stats, df=df)

p_value
```

```
[1] 2.223421e-13
```

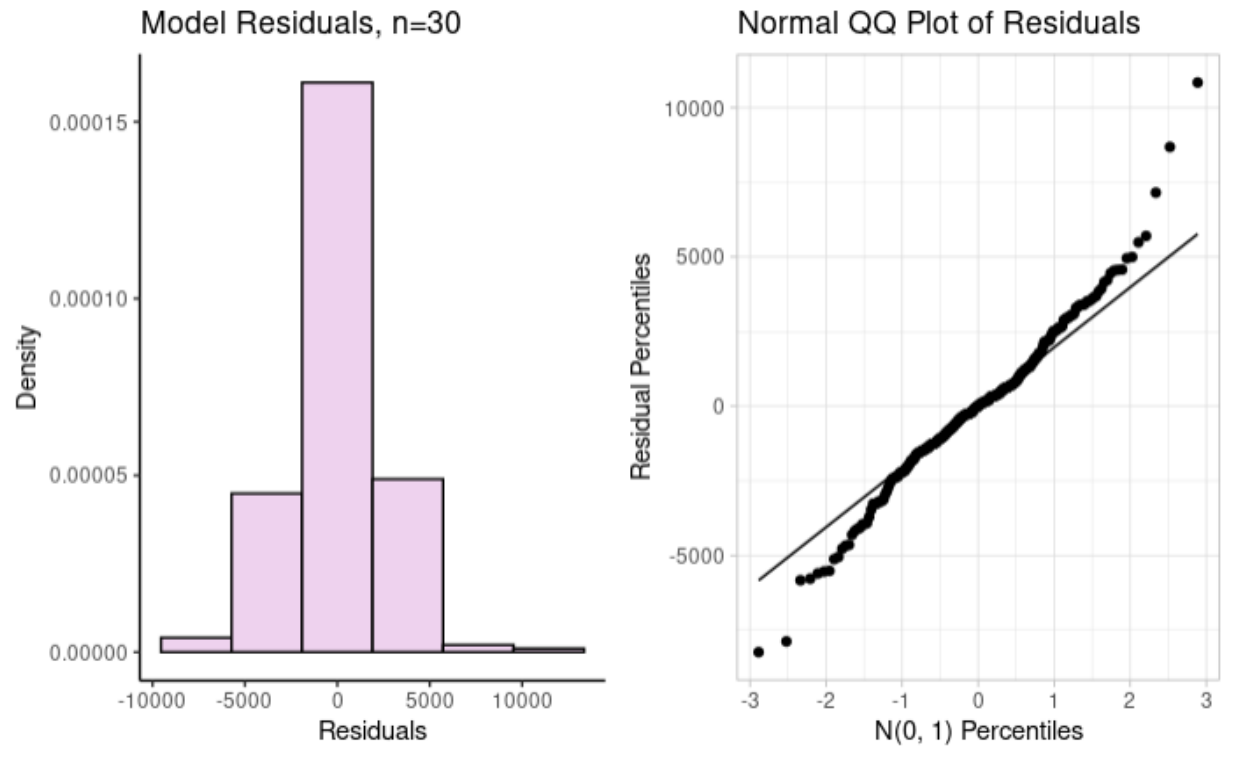
Appendix 2.2

```r
index_attrition <- IBM_clean$Attrition == "Yes"
obs_attrition <- IBM_clean$MonthlyIncome[index_attrition]
obs_not_attrition <- IBM_clean$MonthlyIncome[!index_attrition]

B <- 5000
boot_mean_diff <- c()
set.seed(539)

#Draw bootstrap samples and compute each mean difference in monthly income
for(i in 1:B){
  boot_attrition <- sample(obs_attrition, replace = TRUE)
  boot_not_attrition <- sample(obs_not_attrition, replace = TRUE)
  boot_mean_diff[i] <- mean(boot_not_attrition) - mean(boot_attrition)
}

ci_mean_diff <- quantile(boot_mean_diff, probs = c(0.025, 0.975))
ci_mean_diff
```

```
    2.5%    97.5%
1520.938 2568.749
```

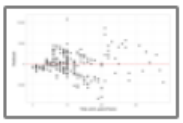
Appendix 3.1

```r
#Checking normality

# Create a new column in SLR.data that will store the residuals
SLR_data$res <- SLR_model$residuals
SLR_data$fit <- SLR_model$fitted.values

#Plot the histogram to observe the distribution of residuals
hist <- ggplot(SLR_data)+
  geom_histogram(aes(x=res, y=..density..),
                 fill='thistle2',
                 colour='black',
                 bins=6)+
  labs(x='Residuals', y='Density',
       title='Model Residuals, n=30')+
  theme_classic()

#Plot the normal QQ plot of residual
qq <- ggplot(SLR_data, aes(sample=res))+
  geom_qq()+
  geom_qq_line()+
  labs(x='N(0, 1) Percentiles',
       y='Residual Percentiles',
       title='Normal QQ Plot of Residuals')+
  theme_light()

grid.arrange(hist, qq, nrow=1)
```

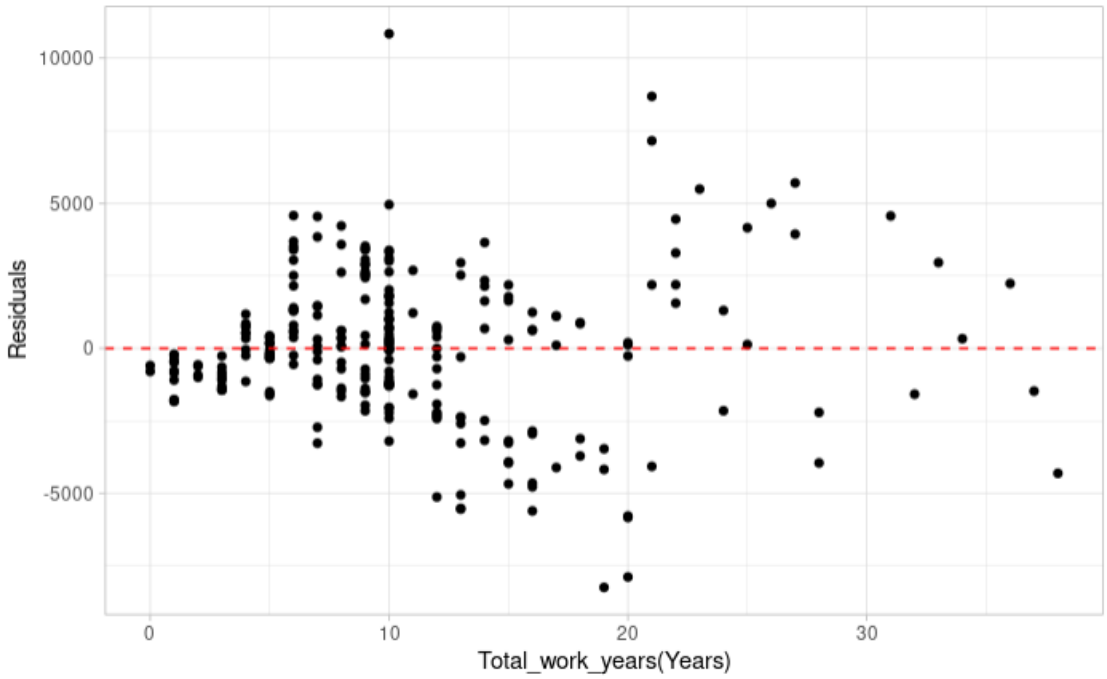


Appendix 3.2

```r
#Plot the residual against total work years
ggplot(SLR_data, aes(x=total_work_year, y=res))+
  geom_point()+
  geom_hline(yintercept=0,
             colour='red',
             lty=2)+
  labs(x='Total_work_years(Years)',
       y='Residuals')+
  theme_light()

#Compute the rounded mean of residuals
round(mean(SLR_data$res),2)
```

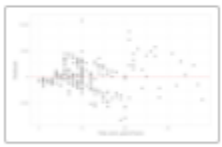


R Console



Appendix 3.3

```r
#Compute the rounded mean of residuals
round(mean(SLR_data$res),2)
```



R Console

```
[1] 0
```

Appendix 4.1

```r
#Plot the scatter plot
ggplot(SLR_data, aes(x=total_work_year, y=monthly_income))+
  geom_point()+
  labs(x = "Total work experience (Years)",
       y = "Monthly Income (Dollars)",
       title = "The scatterplot employee's total work years vs monthly income")
```

Appendix 4.2

```r
SLR_model <- lm(SLR_data$monthly_income ~ SLR_data$total_work_year)

summary(SLR_model)
```

```
Call:
lm(formula = SLR_data$monthly_income ~ SLR_data$total_work_year)

Residuals:
    Min      1Q  Median      3Q     Max
-8238.2 -1388.2    -2.6  1321.7 10840.2

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                2470.11     293.43   8.418 2.79e-15 ***
SLR_data$total_work_year    413.37      22.68  18.227  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2570 on 255 degrees of freedom
Multiple R-squared:  0.5657,    Adjusted R-squared:  0.564
F-statistic: 332.2 on 1 and 255 DF,  p-value: < 2.2e-16
```

Appendix 4.3

```r
#Compute the 95% confidence interval for beta_0 and beta_1

#Compute the critical value, degree of freedom of residual is
#255 from the model summary

t <- qt(c(0.025, 0.975), df=255)

#Compute the 95% Confidence interval for beta_0
CI_beta_0 <- 2470.11 + t*293.43

#Compute the 95% Confidence interval for beta_1
CI_beta_1 <- 413.37 + t*22.68

CI_beta_0
CI_beta_1
```

```
[1] 1892.255 3047.965
[1] 368.706 458.034
```

Bibliography

Jain, R., Shahid, A., Saud, S., & Ramirez, J. (2017). IBM HR Analytics Employee Attrition & Performance. Retrieved March 19, 2022, from https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA_Fn-UseC_-HR-Employee-Attrition.csv.