The Impact of Gender on Income Differences

STAT 302: Accelerated Introduction to Statistics

Team Little Badgers Zhiyi Chen Yushu Wu Diwen Liu

December 20, 2019

Contents

1 Abstract

This study intends to find the relationship between gender and income gaps in the United States. The data in the study is a subsampled merged dataset of multiple datasets from U.S Bureau of Labor Statistics, OECD, U.S Department of Labor, and statistia. The merged dataset from the websites includes information of both female and male labors who once participated or still participating in the labor force in the U.S in the past 20 years. To be more specific, this study samples different number of cases in different tests, as the data is collected from combined datasets of several websites, and the population of these datasets may vary. But, the collected data from original datasets remains representativeness, and the data is used to analyze and infer the possible relationships within all the female and male full-time workers. Moreover, the relation between variables are test by four types of statistical test such as randomization test, t-test, chi-square test, and linear regression. The explanatory variables used in these tests are weekly working hours, major, occupation, age group, and educational level. The respons variables are employment rate, unemployment rate, income, and income level. Through the four different statistical tests, the hypothesis that gender affects income gaps is consistent with the result of the study. There exists an income gap between male and female full-time workers in different occupations. There are siginificant association between educational level and income level, with exception for the relations of employment rate and unemployment rate. Therefore, it is necessary to study gender as a factor contributes to income gaps in the United States.

2 Introduction

2.1 Background Information

Income inequality has always been a serious economic topic since 1970s, when the income gap started to rise until the economic boom in the 1990s. The income gap was considered as the difference between "rich and poor"; however, after 80 years of the evolution of women workforce since the World War II ended, gender has always been a factor when it comes to determine salary in the U.S (O'Brien, 2015, p.1). According to the most recent Bureau of Labor Statistic, among full-time workers, women earn about 78 cent comparing to a man's dollar. Dumontet et al. (2012) conducted a study by saying that the income gap between males and females stil exist nowadays regardless of country, status, and career (p.1). For instance, he suggests that women are having an income 26% less than their male colleagues in France(p.6). Indeed, the gender income gap has already been a worldwide issue that the economics, specialists and governments are putting emphasis on to determine the countries welfare and income per capita, etc.

Zoom in to the United States as a representative, looking back in 2017, the Louisiana stat had the largest gender pay gap with a ratio of 69% between males and females, and California had the lowest gender pay gap with a ratio of 89% (as cited in Vagins, 2019, p.1). Moreover, data collected within 2009 and 2013 by the American Community Survey shows that among the 3.5 million American households, women are usually employed to a lower paid occupation than men (as cited in O'Brien, 2015, p.1). Therefore, it is vital to examine the association between the gender and the income differences. By studying their relationships, there would be a suggestion to reduce the income gap by improving several sectors such as education level, career selections, family policies, and working age.

Hopefully, the results and analyses of this study can provide practical suggestions and usage to the government, economists, and gender-inequality-study specialists. Regulations and improvements could be implemented to reduce the gender income gaps.

2.2 Research Objectives

As the exact aspects of gender differences that are associated with income gaps still remain unclear, and most studies before have their own purpose and focus on examining the income differences between genders in general. This study will pay more attention on analyzing the the practical aspects that lead to gender income gaps. Based on the prior studies conducted by others, the income differences rate between males and females are actually decreasing yearly, but it still remains a problem in the nowadays society. Moreover, the previous studies mainly focus on the income differences in general and take several representatives as examples. Therefore, comparing with the previous studies that focus on several countries or careers, which make the data scatter, the advantage of this study is that the study will mainly focus on the income differences specifically in the United States. The sample will be representative of the general situation such that it will reduce the bias and other confounding variables due to the differences in states, careers and etc.

3 Methods

3.1 Population of interests

The population of interest in this research paper is all male and female full-time workers in the United States.

3.2 Type of Study

The type of study is observational study because the reaserch does not actively control any of the explanatory variables involved (i.e. majors, education level, average working time, occupations and age group). Instead, the datasets used in this research are simply overserved and collected from statistics and data that already existed. This research neither controls explanatory variables nor generates new data. Thus, this study is an observational study.

3.3 Data Collection and Representativeness

The data set used in this research is a merged dataset from multiple datasets from the webites, CNBC, data USA, IWPR, AAUW, and Bureau of Labor Statistic.

The representative of these data is examined. The most data of different majors are derived from AAUW and IWPR. Those data of different occupations are collected from the whole United States. In each occupation, the income of workers is evaluated and been separated into groups of ganders: males and females. This grouping operation can better help our work in estimating the tests. The data of men and women's income in different ages are collected from CNBC, which records men's income and women's income separately within seven different ranges between 16 and 65 or order. The data of the total income in different life ranges is recorded in the website of BLS (Bureau of Labor Statistics) and is recorded more precisely within 12 ranges from 16 years to 65 years and over.

Since the dataset is merged from multiple dataset, some manipulations are required to the data. Since the income separately calculated between ranges is the average of the total income of a gender in all occupations, thus we cannot view theses data the same way as how we exam "separating occupations". We put the data into two groups with one specified on occupations, and the other one specified on ages, and then do the t-test for difference between means separately.

3.4 Sampling

This research will use the stratified random sample and randomly select 1000 people within 5 different majors from United States with each major of size 200. Each group of size 200 would have equal chance to become the sample. This way of sample can effectively prevent most bias and give precise data as much as possible, and could produce a sample that can represents the population of all workers in these five majors in United States. Some bias may still exist, despite simple random sampling is used: some workers were not recorded by their company or factory; the record of workers has not been updated to the newest and still remining in the last year's data. It is hard to change and get the most precise data, but we are able to lower the error to the largest extent by employing the latest records.

3.5 Variables of interest

3.5.1 Explanatory Variables

Names	Types	Level
Weekly Average Working time	Numerical	N.A
Major	Categorical	N.A
Occupations	Categorical	N.A
Age Group	Categorical	[15, 24] [25, 54] [55, 64]
Educational Level	Categorical	High school Bachelor's Degree Master's Degree Doctoral Degree

In explanatory variables, Major is defined as the particular subjects that male or female full-time workers specialized at college or university. The major has impact workers acquired, and some majors are more valued than others, which would be an important factor affecting their careers, so does income. Second, the Education Level is defined as the educational attainment that males and females worker have completed. Education Level can affect the income to a large extent, and it is assumed that higher education level is related to higher income. Third, weekly average working time in hours refers to the total average amount of time that full-time male and female workers spend on works within their career time on a weekly basis. The consistency of working time is important for the comparison of the income between male full-time workers and female full-time workers. As most of the careers evaluate their applicants working experience by determining their hourly working time before, and the working experience will directly affect the income. Then, Occupation is the industries that both women and men are working at, and the study will focus on the mean income differences of different industries. Last, age is defined by the different age groups and is categorized into three groups (age from 15 to 24, 25 to 54, and 55 to 64).

3.5.2 Response Variables

Names	Types	Level
Employment rate	Numerical	[0, 1]
Unemployment rate	Numerical	[0, 1]
Income	Numerical	in dollars
Income Level	Categorical	Low Below Average Above Average High

In response variables, first, the employment rate is the total number of employment of both males and females divided by the whole applicants of the job in different kinds of area. Second, the unemployment rate is the total number of unemployment of both males and females divided by the whole population who are not in the labor force market but are willing to find a job. Third, the income is defined as the salaries or wages that male and female full-time workers received on an annual based work. Last, the income level can be divided into four classes Low, Below Average, Above Average, High according to the average full-time workers' income, which is defined as the average income of both male and female individual full-time workers who work in the U.S labor markets. The income level can reflect the status and also income of female full-time workers, and the income level will be examined with the association of educational level. As a result, these reponse variables summarize the overall income gaps and gender bias in some sorts of degree.

3.6 Statistical Test

This project will use four statistical tests to answer the questions of this project: randomization test, t-test, chi-square test, and Linear Regression. First, randomization test is used to test the hypothesis that male full-time workers' gains higher average income than female full-time workers in different kinds of industries. Second, t-test is used to test the hypothesis that the mean unemployment rate of male full-time workers is smaller than female full-time workers in all age groups. Last, the linear regression is used to test whether there is a correlation between weekly average working time of female full-time workers and female employment rate and find the coefficient of linear regression if the correlation actually exists.

This study uses the R-Studio as the computational software to do the statistical test on the data. The use of statistical tests is achieved through R-Studio and the packages (dplyr, ggplot2, knitr, grid, gridExtra, ggfortify, reshape2, xtable, Lock5Data)

4 Results

4.1 Randomization Test: Occupations and Average Income

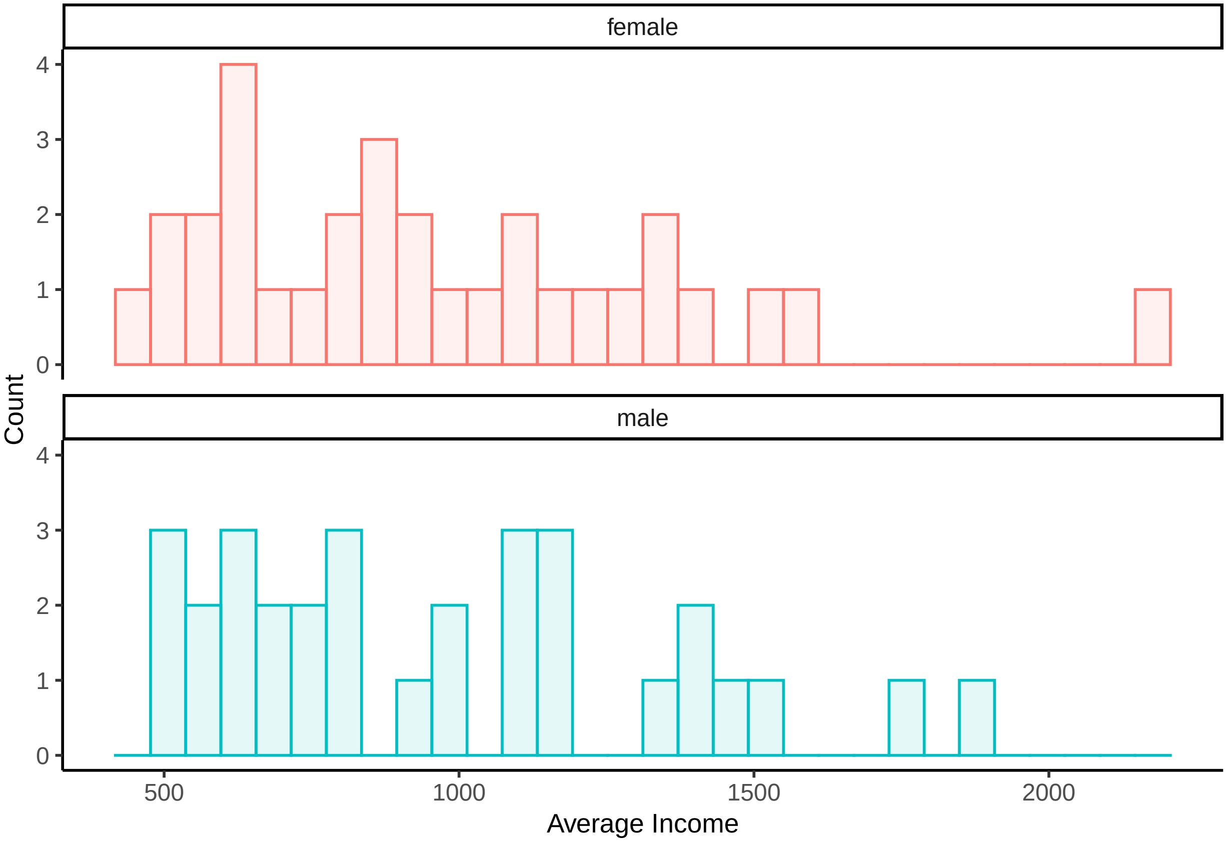
The variables of this test are Occupations, which are defined as different industries that both men and women are working at, and Average Income, which will be defined as the mean weekly wages that both men and women received in different industries, and the industries will include Management, Business and Financial Operations, Professional Jobs, Engineering and Architecture, Physical Science Occupations, Social Service Occupations, Legal Occupations, Education Occupations, etc. The question derived from the research question is that whether male full-time workers have higher average incomes than female full-time workers in different kinds of industries. This test will combine all the average incomes on weekly basis from different occupations. Hence, the aim of this test is to test if male full-time workers have a higher average incomes than female full-time workers in different industries.

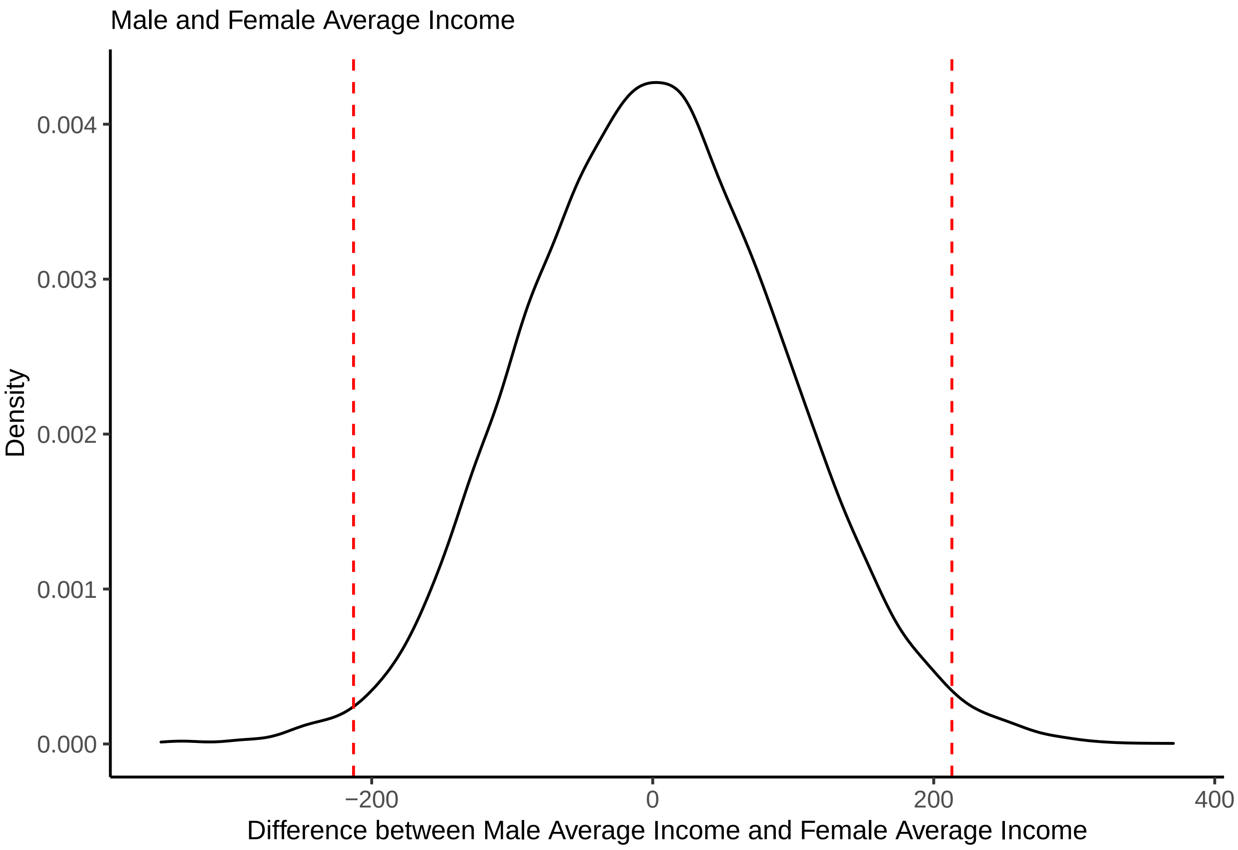
Hypotheses

Define μ_{male} as the average income of male full-time workers on weekly basis and μ_{female} as the average income of female full-time workers on weekly basis.

$$H_0 : \mu_{male} \leq \mu_{female}$$
$$H_a : \mu_{male} > \mu_{female}$$

Summary Figure





```
    2.5%     97.5% 
-174.6169  186.5839 
```

Check for Assumption

Note that the randomization distribution is approximately symmetric and bell-shaped.

Confidence Interval

We are 95% sure that the average income of male and female full-time workers is between the interval of -174.6169 and 186.5839.

Compute p-value

```
the p value with random distribution is 0.0216
```

Interpretation

There is significant evidence such that the mean income of male full-time workers is higher than the mean income of female full-time workers. Therefore, there exists a wage gap in different industries among male full-time workers and female full-time workers. (one side two independent samples randomization test, $\bar{x}_{male} - \bar{x}_{female} = 212.94$, $n_{male} = 31$, $n_{female} = 31$, $p = 0.0216$, $\alpha = 0.05$).

4.2 T-Test: Age and Unemployment Rate

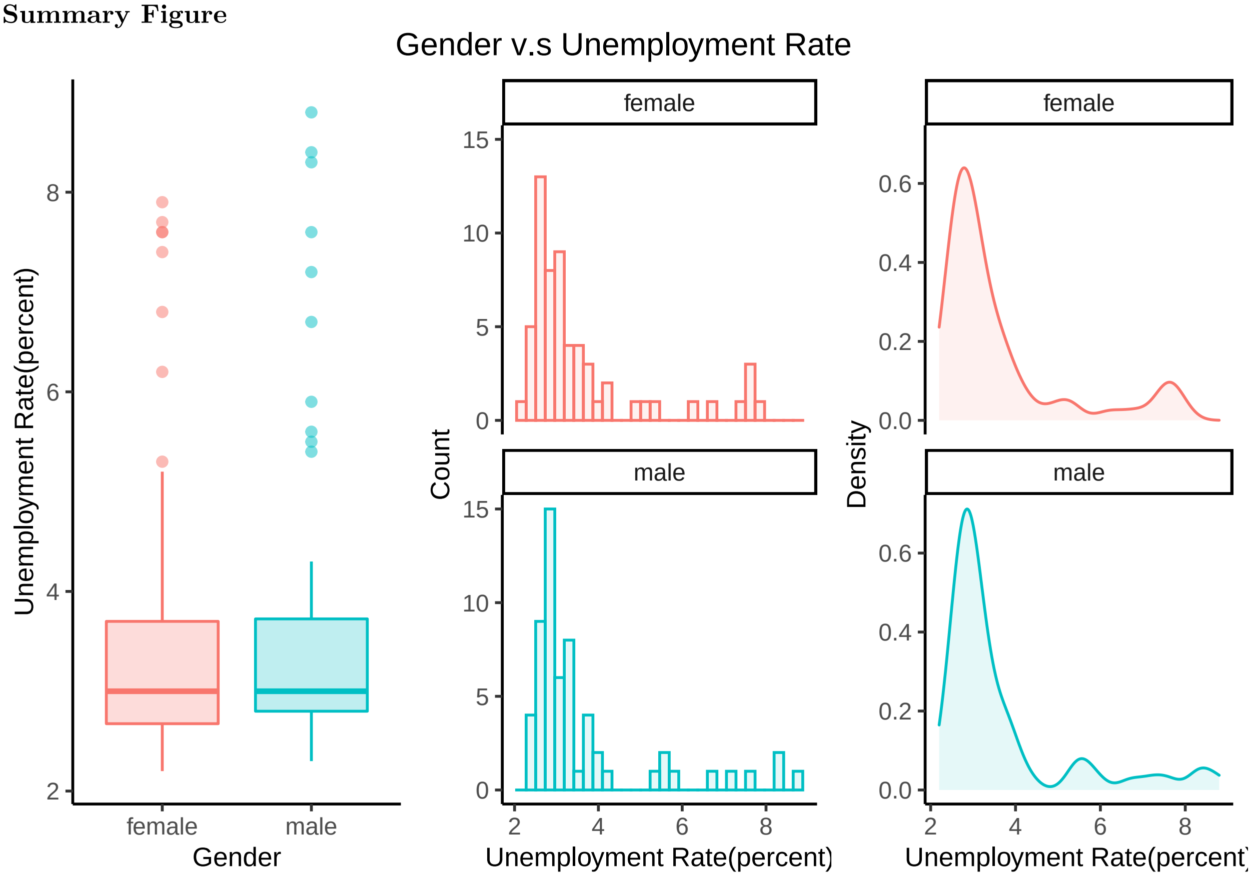
The variables of this test are Age Group, which is defined by the different age groups of both male and female full-time workers, and Unemployment Rate, which, in this case, is population that is not participating in the labor force but are willing to participate. The data will focus on the unemployment rate between two genders among the three age groups (15 to 24, 25 to 54, and 55 to 64) in 2019 in the U.S. The hypothesis is that the mean unemployment rate of male full-time workers are smaller than the mean unemployment rate of female full-time workers. Hence, the aim of this test is to test if the mean unemployment rate of male full-time workers in certain age group is lower than the mean unemployment rate of female full-time workers in the same age range.

Hypotheses

Define μ_{male} as the mean unemployment rate of male full-time workers from all age groups and μ_{female} as the mean unemployment rate of female full-time workers who are also ranged from all age groups.

$$H_0 : \mu_{male} \geq \mu_{female}$$
$$H_a : \mu_{male} < \mu_{female}$$

Summary Figure



Check for Assumption

```
‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

gender	count
female	60
male	60

The sample size is sufficiently large such that n_{male} is bigger than or equal to 30 and n_{female} is bigger than or equal to 30 where n_{male}=60 and n_{female}=60. Therefore, a t-distribution is appropriate.

Calculate Test Statistic

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{3.782 - 3.497}{\sqrt{\frac{1.89^2}{60} + \frac{1.13^2}{60}}} \\ &= 1.003 \end{aligned}$$

Confidence Interval

We are 95% sure that the differences between the mean unemployment rate of male and the mean unemployment rate of female is within the interval -0.284 and 0.853.

Compute P-value

$$\begin{aligned} T &\sim t_{min\{n_1-1,n_2-1\}} \\ p-value &= P(T \geq t) \\ &= P(t_{59} \geq 1.003) \\ &= 0.841 \end{aligned}$$

Intepretation

There is no siginificant evidence that the mean unemployment rate of male full-time workers in all three age groups is smaller than the mean unemployment rate of female full-time workers in all three age groups.(left-tail two sample independent t-test,$\bar{x}_{male} = 3.782$, $\bar{x}_{female} = 3.497$, $n_{male} = 60$, $n_{female} = 60$, $t = 1.003$, $p = 0.841$, $\alpha = 0.05$)

4.3 Chi-Square Test: Educational Level vs Income Level

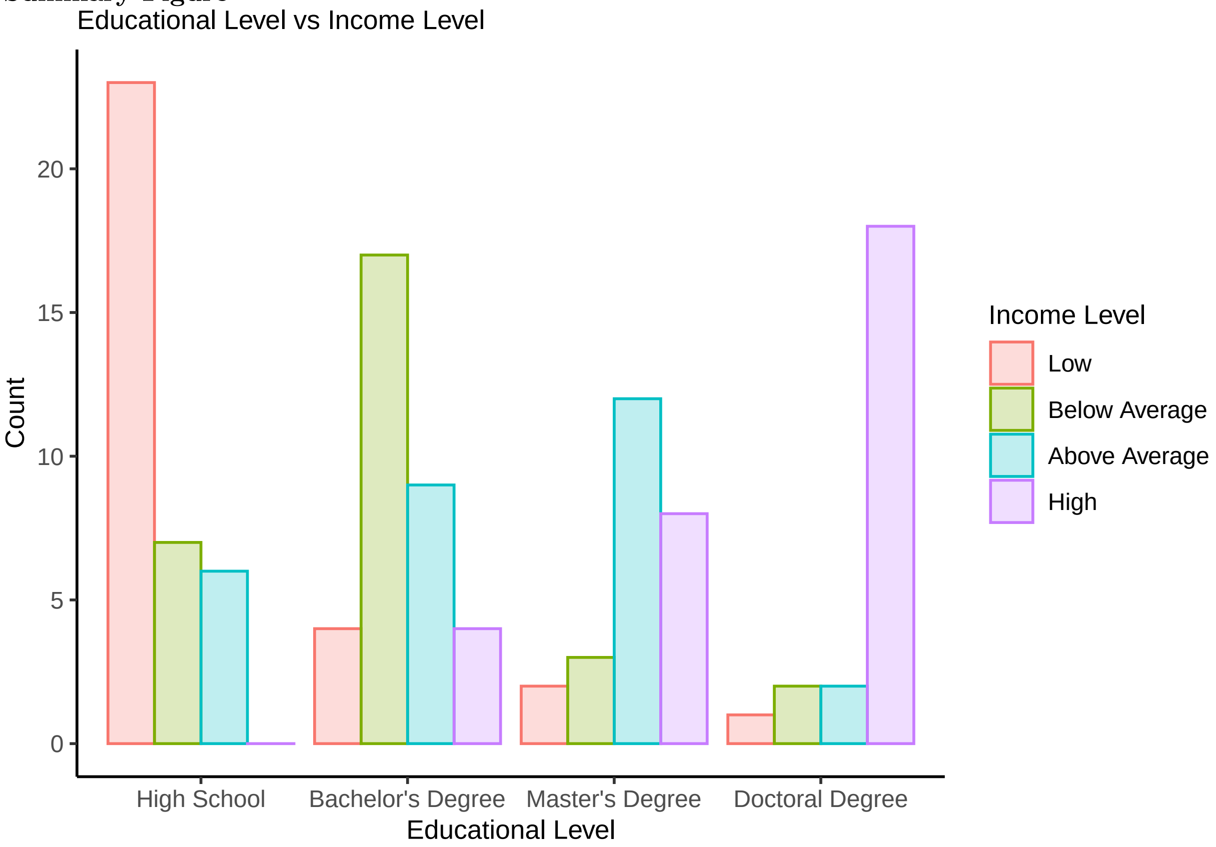
The variables of this test are educational level, which is divided into four classes: High school, Bachelor’s Degree, Master’s Degree, Doctoral Degree, and income level, which is also divided into four classes: Low, Below Average, Above Average, and High. The third test derived from the research question of this project is whether the full-time workers with different educational levels will have different income levels. Hence, this test is to test whether there is an association between the educational level and the income level.

Hypotheses

$$H_0 : \textit{there is no association between educational level and income level.}$$
$$H_a : \textit{there is association between education level and income level.}$$

	Low	Below Average	Above Average	High
High School	23	7	6	0
Bachelor's Degree	4	17	9	4
Master's Degree	2	3	12	8
Doctoral Degree	1	2	2	18

Summary Figure



Check for Assumption

Expected Count Table

	Low	Below Average	Above Average	High
High School	9.152542	8.847458	8.847458	9.152542
Bachelor's Degree	8.644068	8.355932	8.355932	8.644068
Master's Degree	6.355932	6.144068	6.144068	6.355932
Doctoral Degree	5.847458	5.652542	5.652542	5.847458

All the expected counts are 5 or greater. Therefore, it is appropriate to use the chi-square distribution.

Calculate Test Statistic

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{k} \frac{(Observed - Expected)^2}{Expected} \\
&= \frac{(23-9.15)^2}{9.15} + \frac{(7-8.85)^2}{8.85} + \frac{(6-8.85)^2}{8.85} + \frac{(0-9.15)^2}{9.15} + \frac{(4-8.64)^2}{8.64} + \frac{(17-8.36)^2}{8.36} + \frac{(9-8.36)^2}{8.36} + \frac{(4-8.64)^2}{8.64} + (2 \\
&= 89.98305
\end{aligned}
$$

Compute p-value

$$
\begin{aligned}
X^2 &\sim \chi^2_{(r-1)(c-1)} \\
p - value &= P(\chi^2_{(r-1)(c-1)} \geq X^2) \\
&= 1.641 \times 10^{-15}
\end{aligned}
$$

Interpretation

There is significant evidence that the educational level is associated with the income level (chi-square test for association, X^2=89.98305, df=9, p=1.641e-15, α=0.05)

4.4 Linear Regression: Weekly Average Working Time and Employment Rate

The variables of this test are Weekly Average Working Time, which is defined as the average working time of full-time workers spend on their main career on a weekly basis, and Employment Rate,which is defined as the labor force participation rate. This test mainly focus on female full-time workers and is intended to find the correlatoin bewteen female full-time workers weekly average working time and employment rate in the last 20 years in the U.S. The hypothesis in this project is that there should be a correlation between these two variables. And further, it is likely that a higher weekly average working time and more experienced female workers will associate with higher employment rate. Hence, the aim of this test is to test whether there is a correlation between Weekly Average Working Time and Employment Rate of female full-time workers in the past 20 years in the U.S.

Hypotheses

This test tries to model the relation between weekly average working time and employment rate as a linear model defined as follow:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, \ldots, n$$

where X is the explanatory variable weekly average working time, Y is the response variable employment rate, and

$$\varepsilon \sim N(0, \sigma^2)$$

for some standard deviations.

Test if the slope between weekly average working time and employment rate is different from zero.

$$
\begin{aligned}
H_0 &: \beta_1 = 0 \\
H_a &: \beta_1 \neq 0
\end{aligned}
$$

where β_1 is the slope of the least squares line to predict the employment rate based on the weekly average working time of female full-time workers.

Test if the linear association between weekly average working time and employment rate is different from zero.

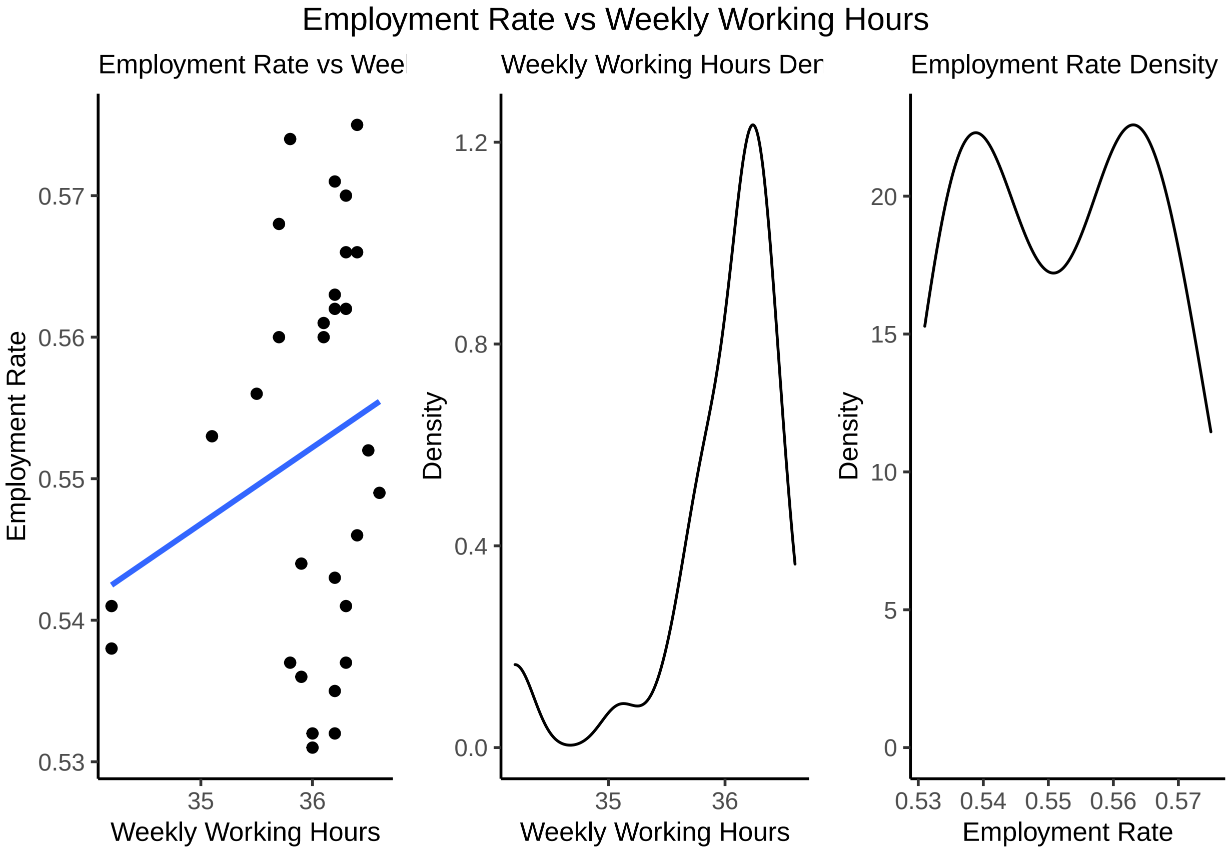
$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0$$

where ρ is the correlation between weekly average working time and employment rate.

These two tests are essentially equivalent.

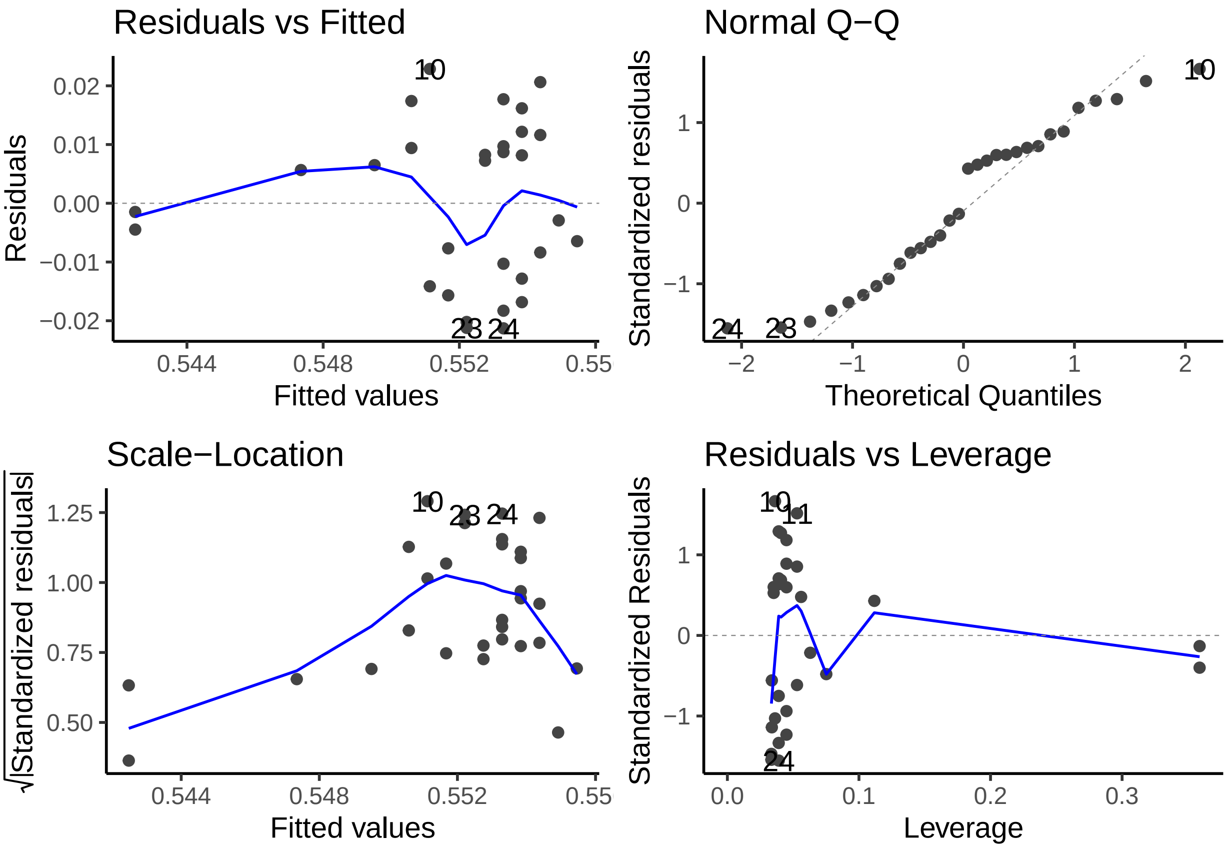
Summary Figures

```
‘geom_smooth()‘ using formula ’y ~ x’
```



Check for Assumptions

```
Warning: ‘arrange_()‘ is deprecated as of dplyr 0.7.0.
Please use ‘arrange()‘ instead.
See vignette(’programming’) for more help
This warning is displayed once every 8 hours.
Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.
```



- The linearity assumption can be checked by the Residuals vs Fitted plot (top-left). There is no obvious fitted pattern via residual plot, the fitted line in blue follows the dashed line, indicating no deviation from linearity, which satisfied assumption.
- The normality assumption can be checked using the Normal Q-Q plot (top-right). The normal probability plot of residuals approximately follows the line y=x although there are some outliers that are far away from y=x. The assumption is satisfied.
- The constant variance assumption can be checked by the Scale-Location plot (bottom-left). There is obvious no fitted pattern via scale-location pattern. However, the trend line is decreasing in the end, but it is not very strong, so the assumption is mostly satisfied.
- The data from each variables does not exactly follow a normal distribution, distribution for weekly working hours skews to the left and the distribution for employment rate slightly skews to the right. However, the sample size is appropriate, so it is eligible to use F-distribution for testing.

Calculate Test Statistic

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3575682	0.1624804	2.200685	0.0361718
weekly.working.hours	0.0054068	0.0045170	1.197000	0.2413429

t* computed by using t-test for the slope between weekly working hours and employment rate.

$$\begin{aligned} t &= \frac{b_1 - 0}{SE} \\ &= \frac{b_1}{SE} \\ &= \frac{0.0054068}{0.0045170} \\ &= 1.196989 \end{aligned}$$

t* computed by using t-test for the correlation between employment rate and weekly working hours.

$$\begin{aligned} t &= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \\ &= \frac{0.2206371\sqrt{30-2}}{\sqrt{1-0.2206371^2}} \\ &= 1.197 \end{aligned}$$

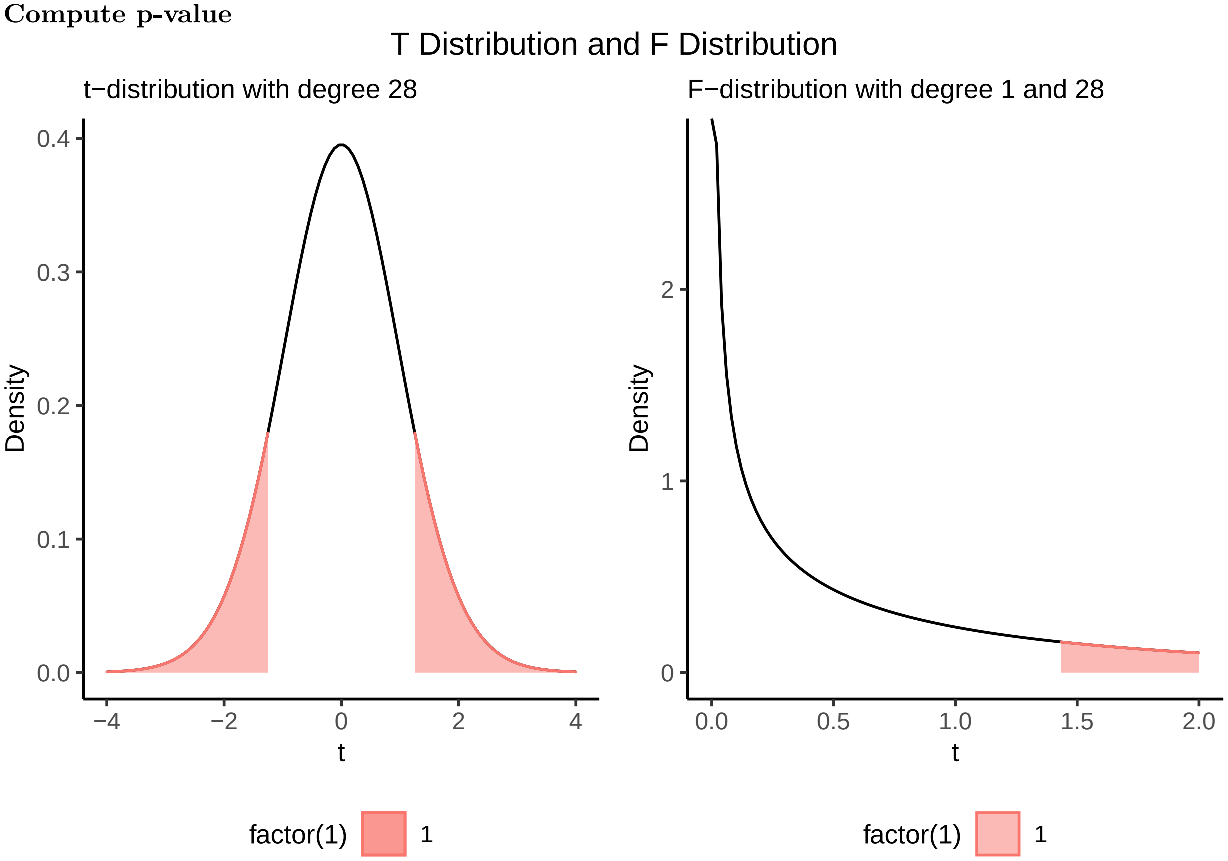
These two t are actually nearly equavilent, only differ by numerical errors

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
weekly.working.hours	1	0.0002803	0.0002803	1.43281	0.2413429
Residuals	28	0.0054767	0.0001956	NA	NA

F-statistic is also computed from ANOVA table

$$\begin{aligned} F &= \frac{MSModel}{MSE} \\ &= \frac{0.0002803}{0.0001956} \\ &= 1.43281 \end{aligned}$$

Compute p-value



For testing t,

$$\begin{aligned} T &\sim t_{n-2} \\ p-value &= P(|T| \geq |t|) \\ &= P(|T| \geq 1.197000) \\ &= 0.2413431 \end{aligned}$$

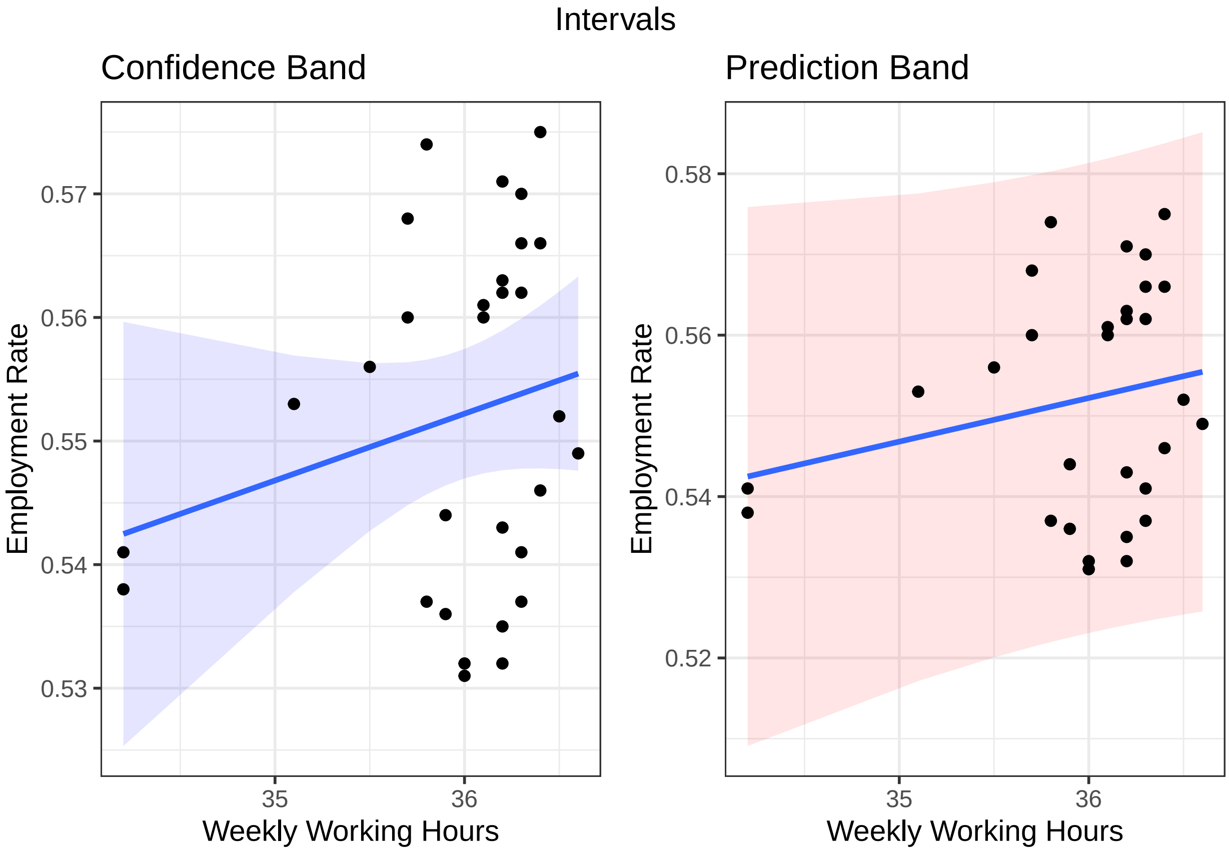
For testing F,

$$\begin{aligned} F &\sim F_{1,n-2} \\ p-value &= P(|F| \geq 1.43281) \\ &= 0.241343 \end{aligned}$$

The p values are equavilent as expected, only differ by numerical error. The p values are represented by the proportion of area shown in the density plots.

Confidence Intervals

```
‘geom_smooth()‘ using formula ’y ~ x’
‘geom_smooth()‘ using formula ’y ~ x’
```



Intepretation

There is no significant evidence that there is a correlation between weekly working hours and employment rate. (two-tail t-test, df=28, p=0.241343, α=0.05)

5 Discussion

The purpose of this study is to analyze whether there is an overall income gap existing in male and female full-time employees in the United States. This research has positive reference value for improving relative regulations and promoting gender equality. In order to reach a conclusive result, several tests have been conducted to decide the association between income and gender of full-time employees in the United States, and the result obtained suggests that there is an income gap between full-time workers of different genders.

5.1 Summary of Findings

The randomization distribution tests the mean differences of average income between male full-time workers and female full-time workers. The occupations and samples are representative and being selected by the BUREAU OF LABOR STATISTICS. Higher average income among different occupations means that male full-time workers earn more wages than female full-time workers in all different occupations. With the wage gap that has been tested, the government should implement policies and reduce taxes to reduce the existing wage gaps. As a result, the occupations are one of the determinators that show income gaps between males and females. Hence, the hypothesis in the project is that the average income of male full-time workers in different occupations is higer than female full-time workers based on the p-value 0.0216. Due to the p-value is smaller than confident level (0.05), the result is statistically significant. Thus the first conclusion would be that male full-time workers earn more incomes than female full-time workers in different kinds of occupations.

The t-test invesitigates the relation between ages and unemployment rate between male full-time workers and female full-time workers. The unemployment represents the economic status of women in the society. Higher unemployment rate means a lower economic status, thus it might come to a suggestion that people with lower economic status would have lower income. Thus, the unemployment rate represents one of the factors to examine the income gap between male and female full-time workers. Another factor age is to examine whether the unemployment rate is associated with age or not. Thus, the hypothesis in the project is that male full-time workers have lower unemployment rate in all age groups (15 to 24, 25 to 54, 55 to 65) than female full-time workers. However, with the test statistic 1.003 and p-value 0.841, the assumption cannot be made as the p-value is bigger than the confident level (0.05), which means the result is not statistically significant. Therefore, male full-time workers do not have a lower unemployment rate than female full-time workers, and age is not a factor that contributes to unemployment rate, which is also not a factor that can be used to examine income gaps.

The chi-square test investigates the association between educational level variable and income level variable. The educational level is defined as the educational attainment that individuals may reach, so the educational level represents the level of education attainment of individual full-time workers, which is an important indicator of income level. Higher educational level reflects higher income level of a full-time worker. Thus, the hypothesis is that there is an association between educational level and income level. From the t-test, the p value (1.641e-15) is less than confident level (0.05), so the result is statistically siginificant. Therefore, it is a strong evidence that there is difference in income level with different educational level. The data can be used in further studies to determine whether females have less number of people obtaining higher educational attainment.

The linear regression examine the correlation between weekly working hours and employment rate, focusing only on female full-time workers. The weekly working hours reflect the time that female full-time workers have spent on works and the work experience they gained when they were spending time on works. The initial assumption is that higher working hours means higher employment rate, which indicates a positive correlation. Hence, as the employment rate is also an indicator of income, the result can be used to determine whether male workers are having a higher employment rate relative to similar working hours, or saying working experience, or relative to less working hours. The hypothesis in the project is that there is a nonzero correlation between weekly working hours and employment rate. However, from the t test, the p-value (0.241343) is much greater than confident level (0.05), so the result is not statistically significant. Therefore, there is no strong evidence that there is a nonzero correlation between these two variables.

In general, the result of the study agrees with the hypothesis that the gender can affect the income gaps. Male full-time workers is proved to earn more than female full-time workers in different kinds of occupations. Also the educational level is associated with income level, where higher educational level implies higher income level, and males are indicated to earn higher degrees in college than females do and more males would earn high degrees than females. However, the study cannot find the assoiation betweem gender and unemployment rate and the association between weekly working hours and employment rate, which undermine the general hypothesis of this study. But, still, the result is representative because the sample selected is representative of the overall circumstances in the U.S. The explanatory variables occupation, major, weekly working hours, educational level, age group are important factors that are representative the income gaps between male and female full-time workers. The response variables average income, income level, employment rate, unemployment rate can directly reflect the difference in incomes within male and female full-time workers. Since, half of the hypothesis test gives a siginificant result, the alternative hypothesis is supported by the analysis. Therefore, those results support the overall hypothesis of this study that the gender has significant influence over income, and the cases that are analyzed in the tests are all contributed to the factor of gender.

5.2 Error Analysis

One possible error is that the income level investigated are not comprehensive enough. This study generally divides the income level of full-time male and female employees to four levels, Low, Below Average, Above Average, and High. However, there might be gaps between these levels, which leads to the omission of certain income groups. This is because the variation of incomes that influence the income level over time in the United States. This study examines the average income of four levels of income, which is not conclusive enough because the income level selected might not be representative, and therefore generates bias. In addition, the dataset which the average working time is calculated from is incomplete. That is to say, the dataset only provides average working time for certain positions, and the calculation is based on those positions offered. As a result, the average working time calculated might be slightly biased. The occupation may also contain bias, as the occupation selected is not conclusive to all the jobs in the U.S. Employment rate and unemployment rate are influenced by many different variaties, so it is hard to find direct relationships between the explanatory variables and employment rate and unemployment rate, so there may exists bias.

5.3 Further Study

Based on the results of this study and former error analysis, one suggestion for the further study would be to invesitgate incomes of more types of occupations in order to increase the representativeness of data and generate a less biased result. For example, occupation can be divided to more detailed categories. In addition, a more complete dataset is suggested because it would improve the representativeness of results.

Besides, more explanatory variable should be examined, including major, which would potentially affect the income. Nevertheless, the datasets for major corresponding to the income groups investigated are currently unavailable. Therefore, the possible effect of major is not disscussed, which should be improved in future studies. Moreover, more tests should be conducted except from randomization test, t-test, chi-square test, and linear regression. For example, the current tests conducted could not conclude a clear relationship between explantory variables and response variabels. Similarly, for the educational level, there are difficulties in revealing its impact on response variables. Furthermore, some data obtained are not symmetric and normally distributed. Therefore, alternative tests might be necessary to examine the validity of results obtained. For example, in linear regression test for weekly average working time and employment rate, data collected for each variavle are skewed and do not follow a normal distribution, which would require further data transformation and alternative tests.

Moreover, the gender distributions in educational levels can be examined to determine the income gaps between female and male full-time workers, as the study only focuses on the association between educational levels and income levels.

6 References

dbplyr_0.8.3

ggplot2_3.2.1

ggfortify_0.4.8

grid_3.6.0

gridExtra_2.3

knitr_1.23

reshape2_1.4.3

xtable_1.8-4

RStudio (version 1.2.1335)

BPS Official. (2015, May 06). Employers prefer male managerial potential to female proven track record. Retrieved November 16, 2019, from https://www.eurekalert.org/pub_releases/2015-05/bps-epm05065.php

Bureau of Labor Statistic: https://www.bls.gov/web/empsit/cpseea10.htm
https://www.bls.gov/opub/reports/womens-earnings/2018/pdf/home.pdf

Dumontet, M., Vaillant, M.L., & Franc, C. (2012). What determines the income gap between French male and female GPs - the role of medical practices. BMC Family Practice, 13(1). doi:10.1186/1471-2296-13-94

OECD: https://stats.oecd.org/Index.aspx?DataSetCode=ANHRS

statista: https://www.statista.com/statistics/192396/employment-rate-of-women-in-the-us-since-1990/

Vagins: https://www.monster.com/career-advice/article/salary-negotiation-gender-wage-gap

7 Appendix

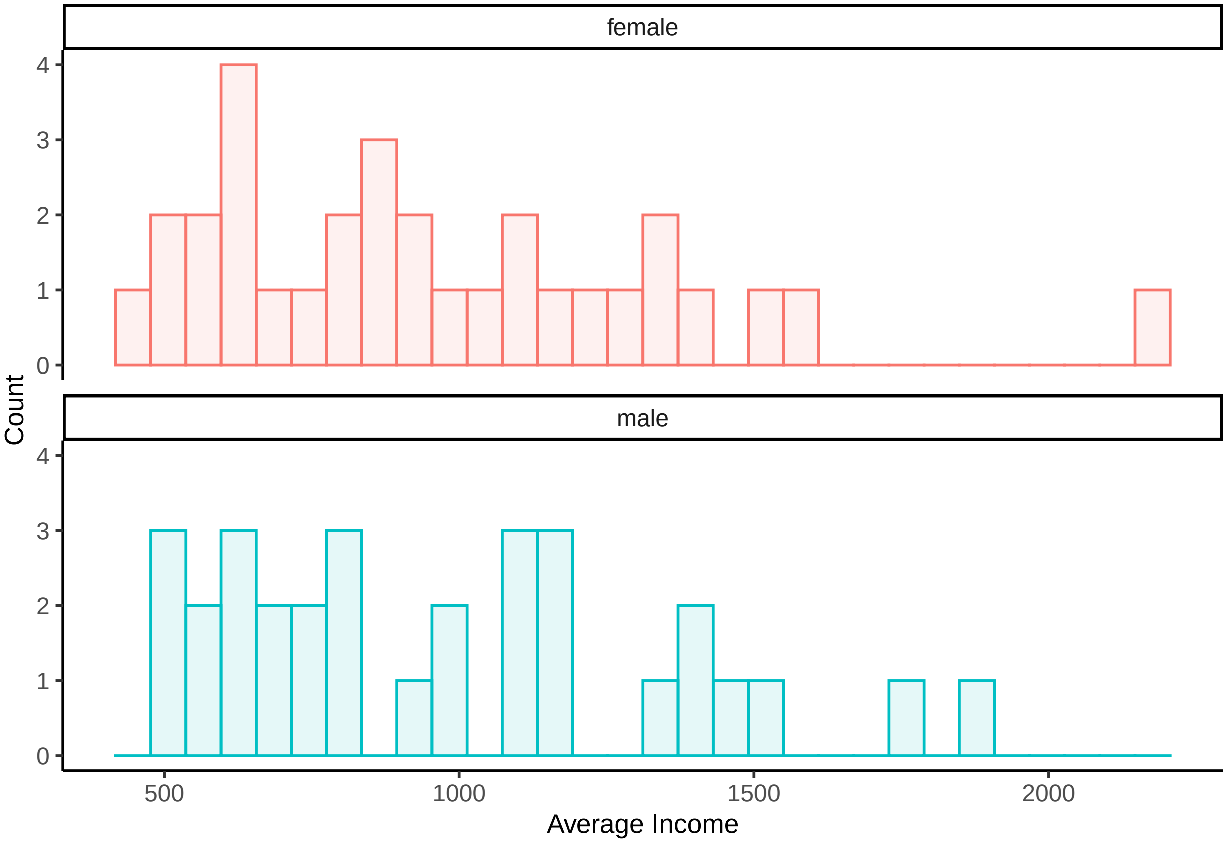
7.1 Packages

```
library(dplyr)
library(ggplot2)
library(knitr)
library(grid)
library(gridExtra)
library(xtable)
library(ggfortify)
library(reshape2)
```

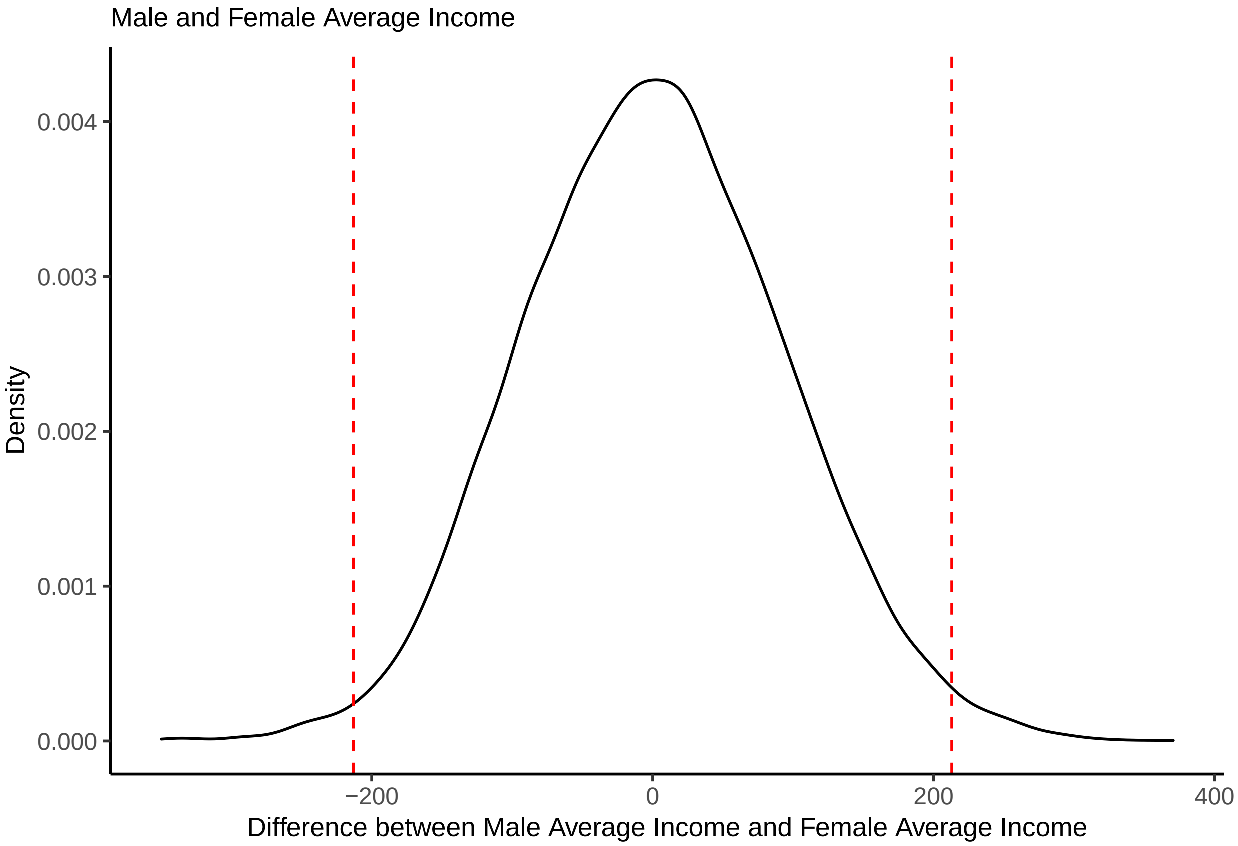
7.2 Results

```


male.income <- c(1468, 1537, 1852, 1383, 1425, 1604, 1528, 1357, 984, 2202, 1148, 1226, 1095, 1151, 1383
female.income <- c(1078, 1168, 1362, 1105, 1024, 1345, 1259, 1156, 886, 1762, 982, 1092, 840, 997, 1078

mean.male.income <- mean(male.income)
mean.female.income <- mean(female.income)
n.male.income <- length(male.income)
n.female.income <- length(female.income)

df.income <- data.frame (
  gender = c("male", "female"),
  average.income = c(male.income, female.income),
  count = c(n.male.income, n.female.income)
)

df.income %>%
  ggplot(aes(x=average.income))+
  geom_histogram(aes(color=gender, fill=gender), bins = 30, boundary = 0, alpha =0.1) +
  facet_wrap(gender~., nrow = 2) +
  labs (x= "Average Income",
         y= "Count") +
  theme_classic()+
  theme(plot.title = element_text(size = 10),
        axis.title = element_text(size = 10),
        legend.title = element_text(size=10)) +
  theme(legend.position = "none")
```

7.2.1 Randomization Test: Randomization Test: Occupations and Average Income



```
set.seed(302)
B<-10000
total <- male.income + female.income
x.bar <- mean(total)
shift.m <- mean.male.income - x.bar
shift.f <- mean.female.income - x.bar

male.income.0 <- male.income - shift.m
female.income.0 <- female.income - shift.f

mat.rand.male <- matrix(sample(male.income.0, B*n.male.income, replace = TRUE),
                        byrow = TRUE,
                        nrow = B,
                        ncol = n.male.income)
mat.rand.female <- matrix(sample(female.income.0, B*n.female.income, replace = TRUE),
                          byrow = TRUE,
                          nrow = B,
                          ncol = n.female.income)

rand.mean.male <- apply (mat.rand.male, 1, mean)
rand.mean.female <- apply (mat.rand.female, 1, mean)
rand.diff <- rand.mean.male - rand.mean.female

df.rand<-data.frame(rand.diff)
ggplot(df.rand, aes(x=rand.diff))+
  geom_density() +
  geom_vline(xintercept = mean.male.income-mean.female.income, color="red", linetype="dashed") +
  geom_vline(xintercept = -(mean.male.income-mean.female.income), color="red", linetype = "dashed") +
  labs(x="Difference between Male Average Income and Female Average Income",
       y="Density",
       title = "Male and Female Average Income") +
  theme_classic() +
  theme(plot.title = element_text(size=10),
        axis.title = element_text(size=10),
        legend.title = element_text(size=10))
```



```
#### Confidence Interval
(ci.per.95 <- quantile(rand.diff, c(0.025, 0.975)))
```

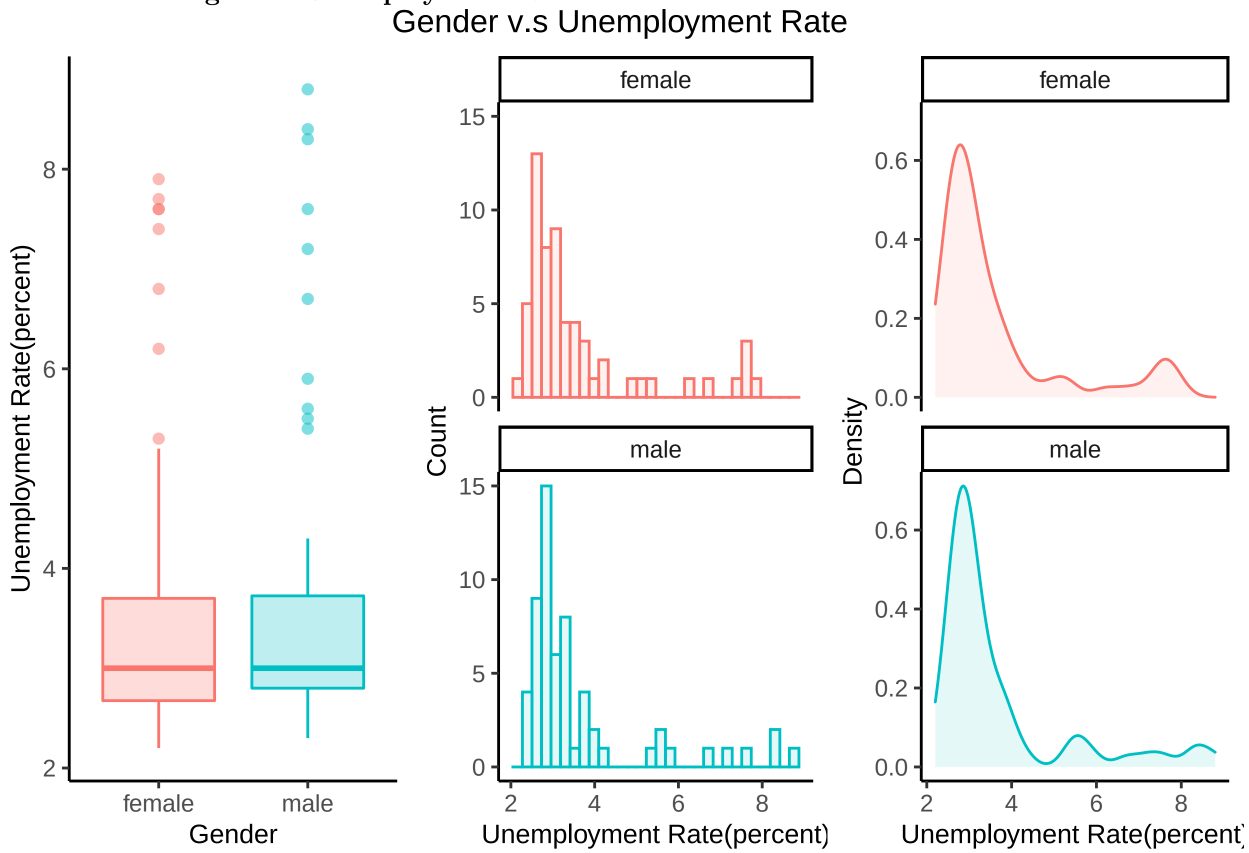
```
##      2.5%     97.5%
## -174.6169  186.5839
```

```
#### Compute p-value
tol <- 1.0e-12
p.value <- mean (abs(rand.diff)>= abs (mean.male.income-mean.female.income)-tol)
cat(sep = "", "the p value with random distribution is ", p.value, "\n")
```

```
## the p value with random distribution is 0.0216
```

```
male <- c(8.4, 7.6, 8.8, 7.7, 8.3, 7.4, 7.6, 7.9, 7.2, 7.6, 3.3, 3.2, 3.2, 3.1, 2.9, 3.0, 3.1, 3.1, 3.0
female <- c(6.7, 6.8, 5.5, 5.3, 5.6, 5.2, 5.9, 6.2, 5.4, 4.8, 3.3, 3.2, 3.3, 3.0, 2.9, 3.1, 3.2, 3.1, 3
x.bar.male <- mean(male)
x.bar.female <- mean(female)
s.male <- sd(male)
s.female <- sd(female)
n.male <- length(male)
n.female <- length(female)
se <- sqrt((s.male^2/n.male)+(s.female^2/n.female))
t<-(x.bar.male-x.bar.female)/se

conf.level <- 0.95
alpha <- 1-conf.level
t.star <- qt(1-alpha/2,min(n.male-1,n.female-1))
ci<-x.bar.male-x.bar.female + c(-t.star,t.star)*se

test <- t.test(male,female,alternative = "less", conf.level = 0.95)

df <- data.frame(
  gender=c("male", "female"),
  unemployment.rate=c(male,female),
  count = c(n.male, n.female)
  )
p1 = ggplot(df, aes(x=gender, y=unemployment.rate)) +
  geom_boxplot(aes(color=gender, fill=gender),
               alpha=0.25, outlier.alpha = 0.5, outlier.shape = 19, outlier.size = 1.5)+
  labs(x="Gender",
       y="Unemployment Rate(percent)")+
  theme_classic()+
  theme(plot.title = element_text(size=10),
        axis.title = element_text(size=10),
        legend.title = element_text(size=10))+
  theme(legend.position = "none")

p2 <- df %>%
  ggplot(aes(x=unemployment.rate)) +
  geom_histogram(aes(color=gender, fill=gender), bins = 30, boundary = 0, alpha = 0.1)+
  facet_wrap(gender~., nrow=2) +
  labs(x= "Unemployment Rate(percent)",
       y= "Count") +
  theme_classic() +
  theme(plot.title = element_text(size=10),
        axis.title = element_text(size=10),
        legend.title = element_text(size=10)) +
    theme(legend.position = "none")

p3 <- df %>%
  ggplot(aes(x=unemployment.rate))+
  geom_density(aes(color=gender, fill=gender), alpha=0.1)+
  facet_wrap(gender~., nrow=2) +
  labs(x="Unemployment Rate(percent)",
       y="Density") +
  theme_classic() +
  theme(plot.title = element_text(size=10),
        axis.title = element_text(size=10),
        legend.title = element_text(size=10)) +
  theme(legend.position = "none")

grid.arrange(p1, p2, p3, ncol=3, nrow=1, top=textGrob("Gender v.s Unemployment Rate"))
```

7.2.2 T-Test: Age and Unemployment Rate



```
#### Confidence Interval
conf.level <- 0.95
alpha <- 1-conf.level
t.star <- qt(1-alpha/2,min(n.male-1,n.female-1))
ci<-x.bar.male-x.bar.female + c(-t.star,t.star)*se

#### Check for Assumption
t = df%>%group_by(gender) %>%
  summarise(count=n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
kable(t)
```

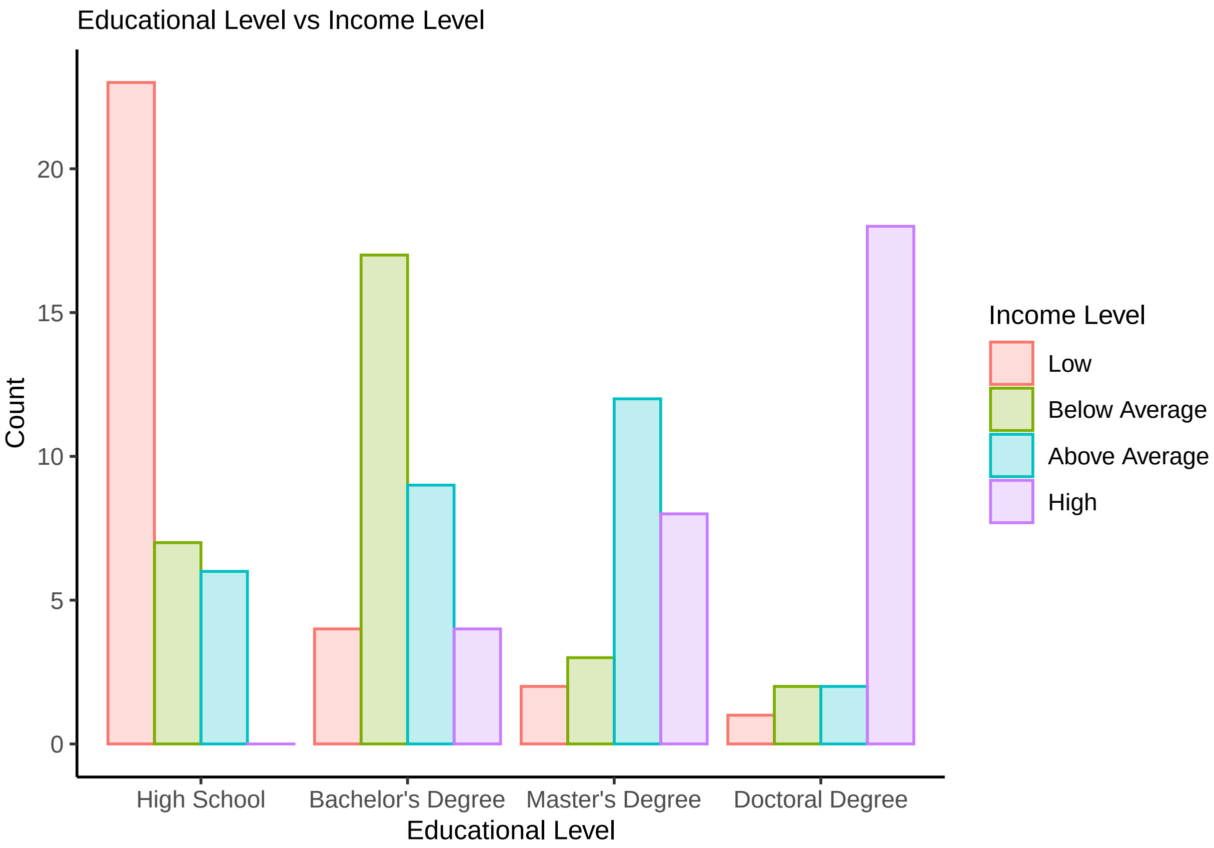
gender	count
female	60
male	60

```
weekly.average.income <- c(1468, 1537, 1852, 1383, 1425, 1604, 1528, 1357, 984, 2202, 1148, 1226, 1095,
educational.level <- c(1, 2, 2, 1, 1, 2, 2, 1, 0, 2, 2, 1, 1, 1, 1, -1, -1, -1, -1, -1, -1, 0, 0, 0, 0,
df <- data.frame(
  income.mean=c(weekly.average.income),
  educational.level=c(educational.level),
  count = c(length(weekly.average.income), length(educational.level))
  )

a = quantile(df$income.mean, 0.25, na.rm = TRUE)
b = quantile(df$income.mean, 0.5, na.rm = TRUE)
c = quantile(df$income.mean, 0.75, na.rm = TRUE)
df = df %>% mutate(income.mean.cut = cut(income.mean, breaks = c(0, a, b, c, 2500), labels = c("Low", "

df = df %>% mutate(educational.level.cut = cut(educational.level, breaks = c(-2, -1, 0, 1, 2), labels =

t = table(df$educational.level.cut, df$income.mean.cut)
dimnames(t) <- list(education = c("High School", "Bachelor's Degree", "Master's Degree", "Doctoral Degr

kable(t)
```

7.2.3 Chi-square Test: Educational Level and Income Level

	Low	Below Average	Above Average	High
High School	23	7	6	0
Bachelor's Degree	4	17	9	4
Master's Degree	2	3	12	8
Doctoral Degree	1	2	2	18

```
#### Summary Figure
melt(t) %>%
  ggplot(aes(x=education, y=value, fill=income)) +
  geom_bar(aes(color=income), position = "dodge", stat = "identity", alpha = 0.25) +
  labs(x = "Educational Level",
       y = "Count",
       title = "Educational Level vs Income Level",
       fill = "Income Level",
       color = "Income Level") +
  theme_classic() +
  theme(plot.title = element_text(size=10),
        axis.title = element_text(size=10),
        legend.title = element_text(size=10))
```



```
#### Check for Assumption
obs.counts <- t

r <- nrow(obs.counts)
c <- ncol(obs.counts)
row.sums <- rowSums(obs.counts)
col.sums <- colSums(obs.counts)

n <- sum(obs.counts)

exp.counts <- outer(row.sums, col.sums, "*")/n
kable(exp.counts)
```

	Low	Below Average	Above Average	High
High School	9.152542	8.847458	8.847458	9.152542
Bachelor's Degree	8.644068	8.355932	8.355932	8.644068
Master's Degree	6.355932	6.144068	6.144068	6.355932
Doctoral Degree	5.847458	5.652542	5.652542	5.847458

```
#### Test Statistic
X.sq <- sum((obs.counts-exp.counts)^2/exp.counts)
X.sq
```

```
## [1] 89.98305
```

```r
#### Compute p-value
pchisq(X.sq, (r-1)*(c-1), lower.tail = FALSE)
```

```
## [1] 1.640873e-15
```

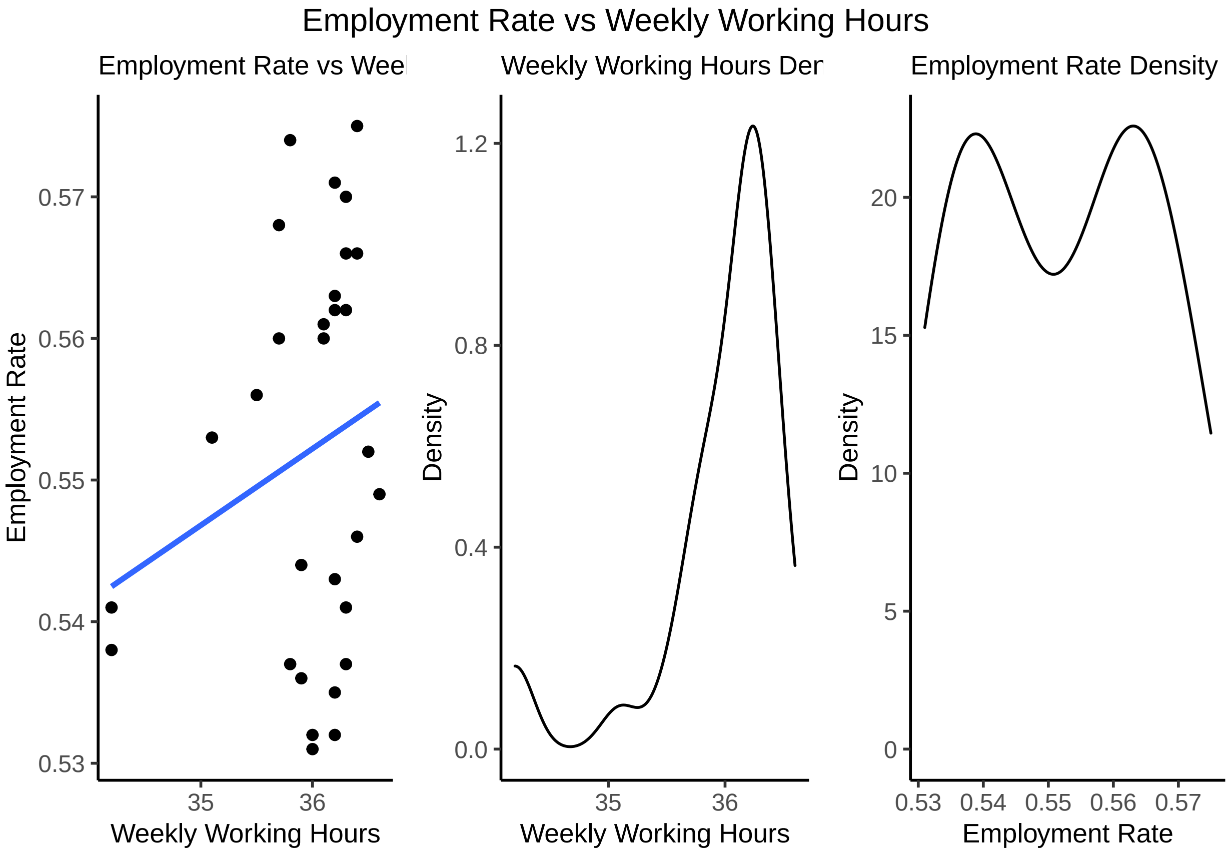
```r
chisq.test(t, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 89.983, df = 9, p-value = 1.641e-15
```

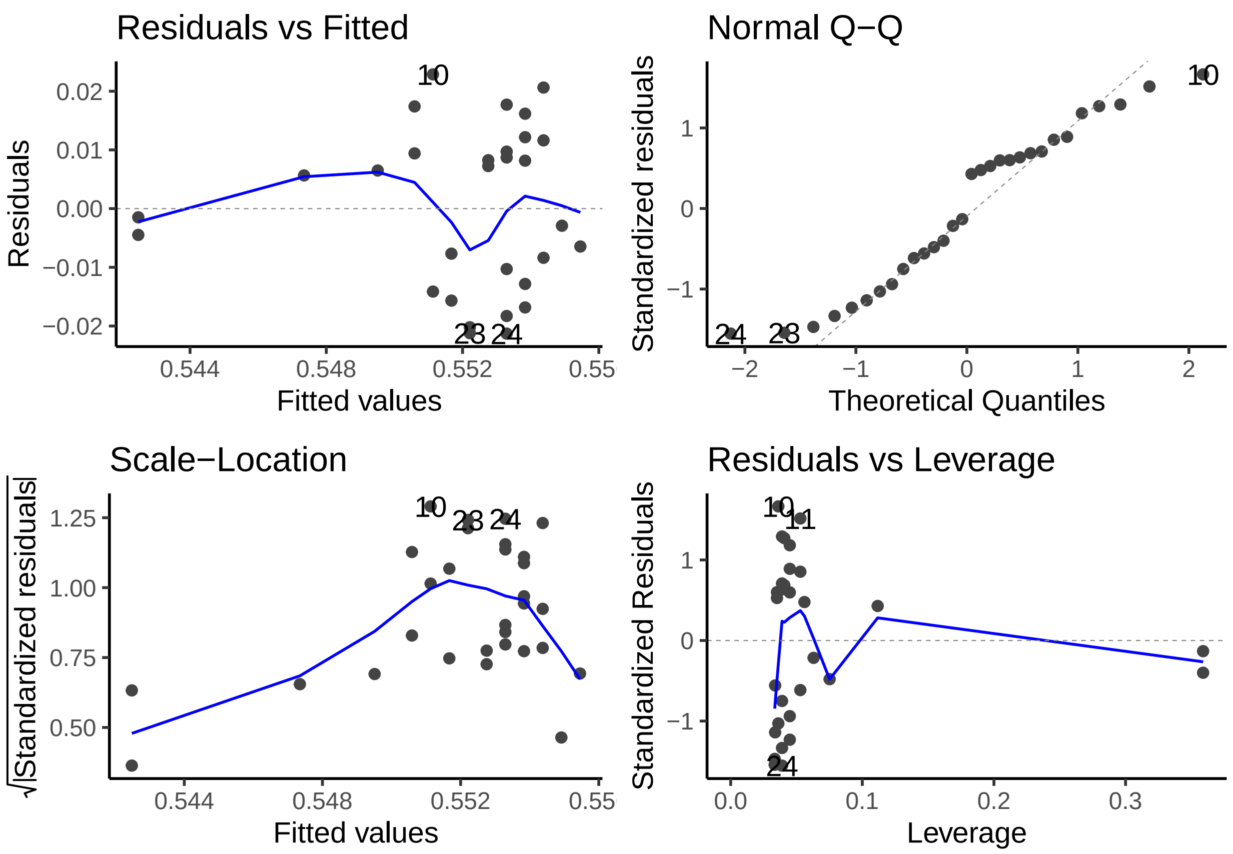
```r
female.weekly.working.hours <- c(36.2, 35.8, 34.2, 34.2, 35.1, 35.5, 35.7, 35.7, 36.2, 35.8, 36.4, 36.3
female.employment.rate <- c(0.543, 0.537, 0.538, 0.541, 0.553, 0.556, 0.56, 0.568, 0.571, 0.574, 0.575,

df <- data.frame(
  employment.rate=c(female.employment.rate),
  weekly.working.hours=c(female.weekly.working.hours),
  count = c(length(female.weekly.working.hours), length(female.employment.rate))
  )

p1 = ggplot(df, aes(x = weekly.working.hours, y = employment.rate)) + geom_point() +
  labs(x = "Weekly Working Hours",
       y = "Employment Rate",
       title = "Employment Rate vs Weekly Working Hours") +
  geom_smooth(method="lm", se=FALSE) +
  theme_classic()+
  theme(plot.title = element_text(size=10),
        axis.title = element_text(size=10),
        legend.title = element_text(size=10))

p2 = ggplot(df, aes(x = weekly.working.hours)) + geom_density() +
  labs(x = "Weekly Working Hours",
       y = "Density",
       title = "Weekly Working Hours Density") +
  theme_classic() +
  theme(plot.title = element_text(size=10),
        axis.title = element_text(size=10),
        legend.title = element_text(size=10))

p3 = ggplot(df, aes(x = employment.rate)) + geom_density() +
  labs(x = "Employment Rate",
       y = "Density",
       title = "Employment Rate Density") +
  theme_classic() +
  theme(plot.title = element_text(size=10),
        axis.title = element_text(size=10),
        legend.title = element_text(size=10))
grid.arrange(p1, p2, p3, ncol = 3, nrow = 1, top = textGrob("Employment Rate vs Weekly Working Hours"))
```

7.2.4 Linear Regression: Weekly Average Working Time and Employment Rate

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```r
#### Check for Assumption
lm.fit <- lm(employment.rate ~ weekly.working.hours, df)
autoplot(lm.fit) + theme_classic()
```



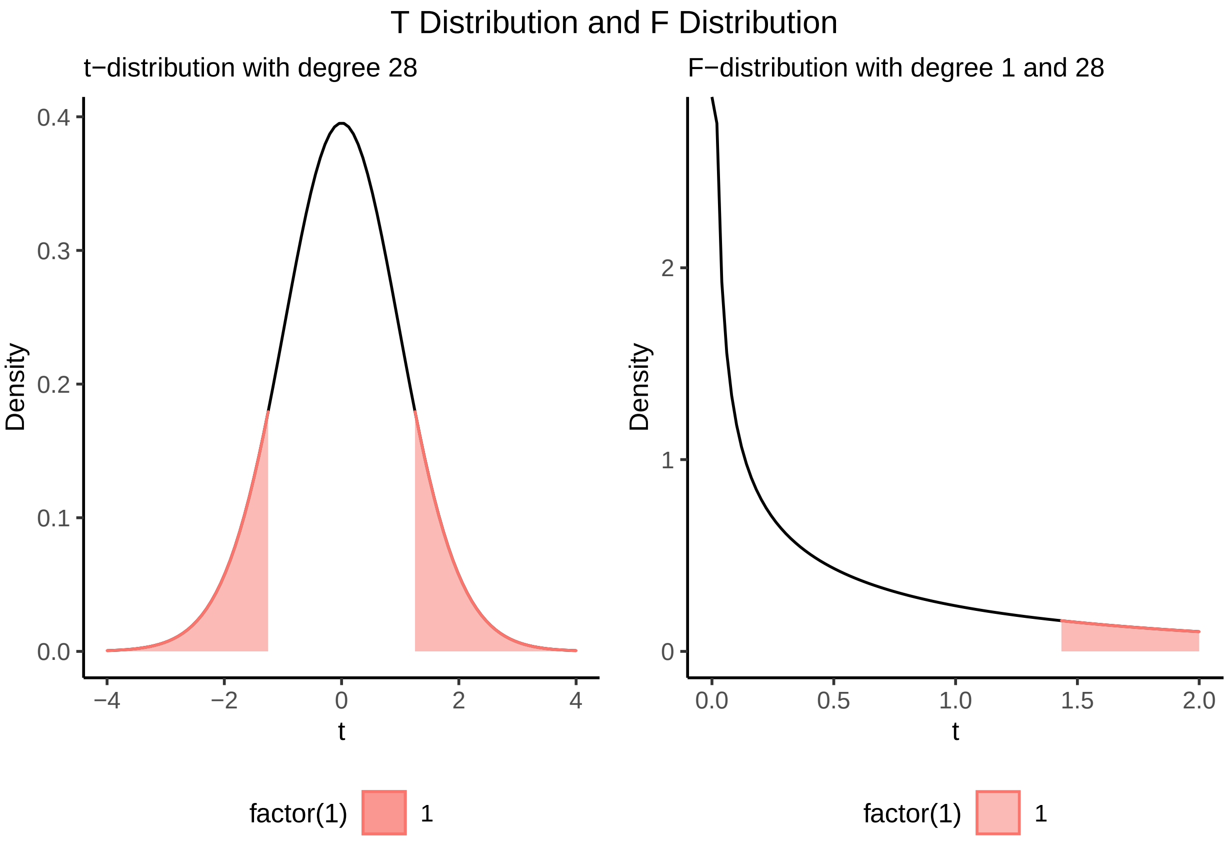
```
#### Calculate Test Statistic
xtable(summary(lm.fit)) %>% kable()
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3575682	0.1624804	2.200685	0.0361718
weekly.working.hours	0.0054068	0.0045170	1.197000	0.2413429

```
r <- cor(df$weekly.working.hours, df$employment.rate)

xtable(anova(lm.fit)) %>% kable()
```

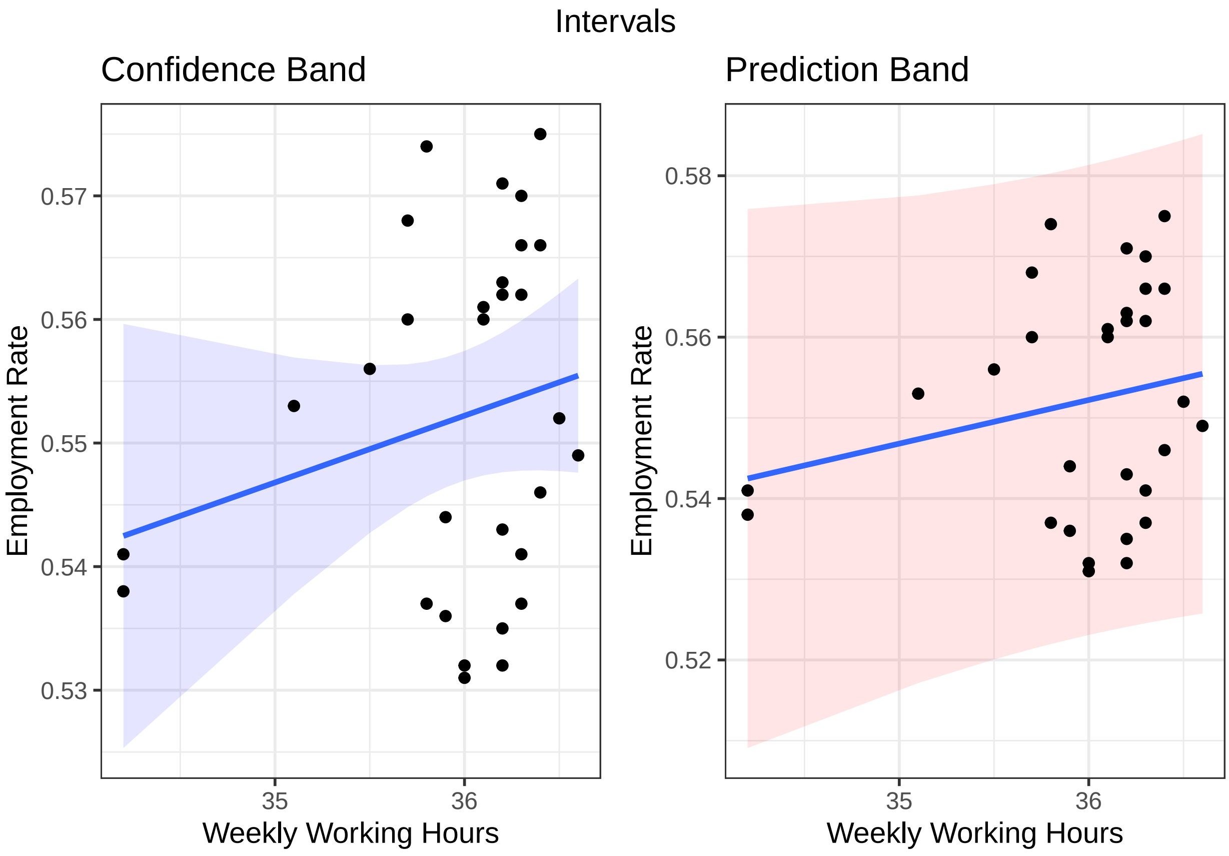
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
weekly.working.hours	1	0.0002803	0.0002803	1.43281	0.2413429
Residuals	28	0.0054767	0.0001956	NA	NA

```
#### Compute P-value
t = 1.197000
x <- seq(-4, 4, length = 100)
y <- dt(x, 28)

gg1 = ggplot(data.frame(x, y), aes(x=x, y=y)) + geom_line() +
  labs(x = "t",
       y = "Density",
       title = "t-distribution with degree 28") +
  theme_classic() +
  theme(plot.title = element_text(size=10),
        axis.title = element_text(size=10),
        legend.title = element_text(size=10))

gg1.data <- ggplot_build(gg1)$data
temp1 <- gg1.data[[1]] %>% filter(x <= -t)
temp2 <- gg1.data[[1]] %>% filter(x >= t)

p1 = gg1 + geom_area(data = temp1, aes(x=x, y=y, color=factor(1), fill=factor(1)), alpha=0.5) +
  geom_area(data = temp2, aes(x=x, y=y, color=factor(1), fill=factor(1)), alpha = 0.5) +
  theme(legend.position = "bottom")

f = 1.43281
x <- seq(0, 2, length=100)
y <- df(x, df1=1, df2=28)

gg2 = ggplot(data.frame(x, y), aes(x=x, y=y)) + geom_line() +
  labs(x = "t",
       y = "Density",
       title = "F-distribution with degree 1 and 28") +
  theme_classic() +
  theme(plot.title = element_text(size=10),
        axis.title = element_text(size=10),
        legend.title = element_text(size=10))

gg2.data <- ggplot_build(gg2)$data
temp <- gg2.data[[1]] %>% filter(x >= f)

p2 = gg2 + geom_area(data=temp, aes(x=x, y=y, color=factor(1), fill=factor(1)), alpha = 0.5) +
  theme(legend.position = "bottom")

grid.arrange(p1, p2, ncol=2, nrow=1, top = textGrob("T Distribution and F Distribution"))
```



```
p.value.t <- pt(1.197000, 28, lower.tail = FALSE) + (1 - pt(1.197000, 28, lower.tail = TRUE))
p.value.f <- pf(1.43281, df1 = 1, df2 = 28, lower.tail = FALSE)

#### Confidence Intervals
conf <- predict(lm.fit, df, interval="confidence")
conf <- data.frame(conf)
df <- df %>%
  mutate(lwr.ci = conf$lwr, upr.ci = conf$upr)

p1 = ggplot(df, aes(x=weekly.working.hours, y=employment.rate, ymin=lwr.ci, ymax=upr.ci)) +
  geom_ribbon(fill="blue", alpha = 0.1) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  xlab("Weekly Working Hours") +
  ylab("Employment Rate") +
  ggtitle("Confidence Band") +
  theme_bw()

pred <- predict (lm.fit, df, interval = "predict")
pred <- data.frame(pred)
df <- df %>%
  mutate(lwr.pred = pred$lwr, upr.pred = pred$upr)
p2 = ggplot(df, aes(x=weekly.working.hours, y=employment.rate, ymin=lwr.pred, ymax=upr.pred)) +
  geom_ribbon(fill="red", alpha=0.1) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  xlab("Weekly Working Hours") +
  ylab("Employment Rate") +
  ggtitle("Prediction Band") +
  theme_bw()
grid.arrange(p1, p2, ncol = 2, nrow = 1, top = textGrob("Intervals"))
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
## ‘geom_smooth()‘ using formula ’y ~ x’
```


Intervals
Confidence Band
Prediction Band
Employment Rate
Weekly Working Hours