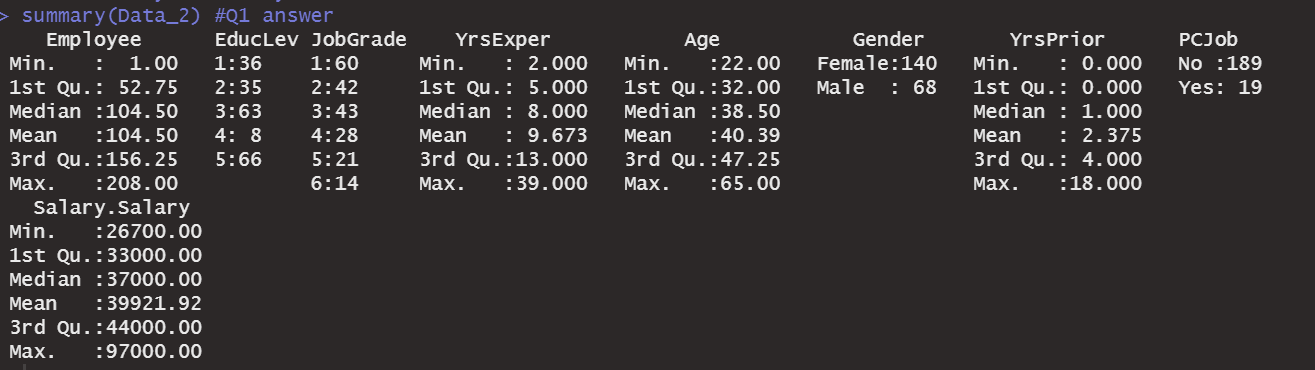
Name: Yu Shiuan Shr

Q1:



**Q2) A plaintiff’s lawyer claims that there is a significant difference in average salary between female employees and male employees. As an analyst for the plaintiff, how would you support this claim? Use a t-test and explain the results as well as your interpretation.**

Assume that the population is the hired employee of this company in the future. And the sample is the current staff list as provided. In this condition, the sigma of population is unknown, and we have 208 (>50) samples as sample size. Even though the data is highly right skewness, the sample of the data will follow normal distribution due to the central limit theorem.

Most of time, we will ensure it follow normal distribution before using T test via Shapiro-Wilk test.

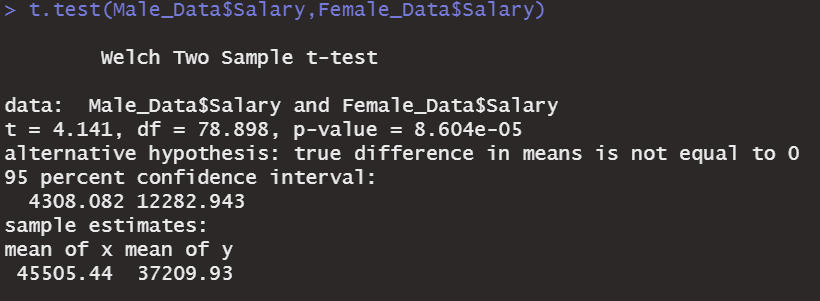
However, due to the requirement of the Q1, we use the t test directly, and I'll provide the result of Shapiro-Wilk and the related normality testing afterwards in case that the prerequisite condition I set above is not correct.

**T- test:**

Let’s assume that random variable X represent the salary. X1 represents the salary of male employee, and the X2 represent the salary of female employee.

H0: μ(X**1)=** μ(X**2)**

H1: μ(X**1) !=** μ(X**2)**



**Details of T- test result:**

Conclusion: p value <0.05. Hence, we reject H0, accept H1.

Therefore, the mean of male salary is not equal to the mean of female salary as the α = 0.05, which supports plaintiff’s lawyer claims

**Supplement part-------------------------------------**

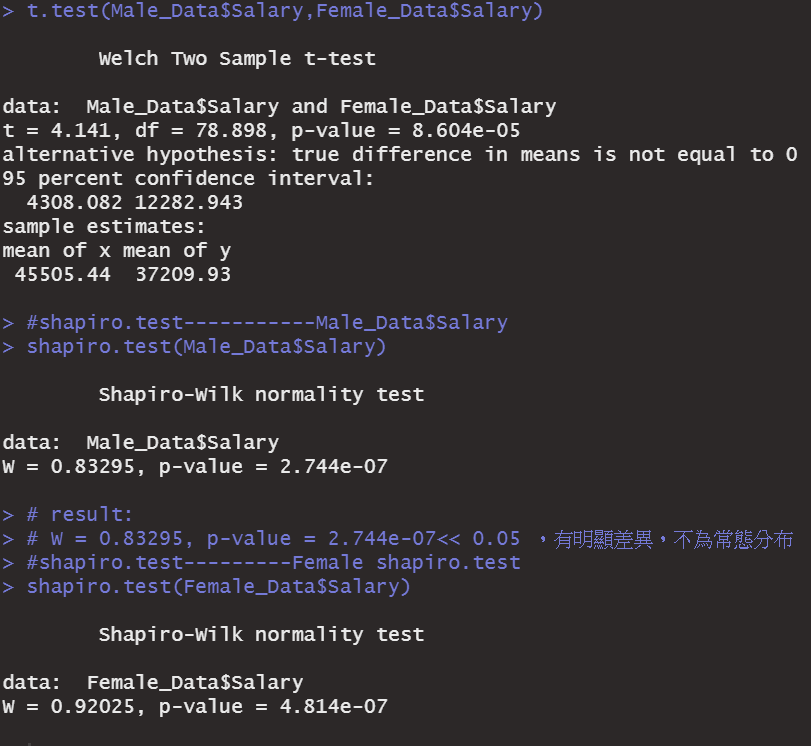
**Test for normality - Shapiro-Wilk:**

Here we provide the Shapiro-Wilk, which supposed to test earlier than t-test.

Let

H0: The sample data are not significantly different than a normal population.

H1 : The sample data are significantly different than a normal population



Details of **Shapiro-Wilk result:**

**Male:**

p-value = 2.744e-07<< 0.05 ，reject H0，it represents it doesn't follow Normal Distribution.

**Female:**

p-value = 4.814e-07 < 0.05，reject H0，which represents that it doesn't follow Normal Distribution.

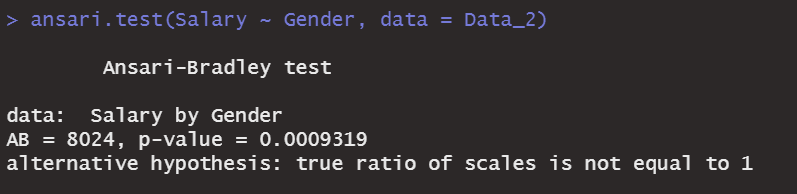
Obviously, both subset’s p- value are far below 0.05, therefore, we reject H0, and accept H1. Which means that the sample data are significantly different from normal population. Therefore, on the condition of comparison of two variable, we choose Ansari-Bardley test to test the variance.

**---------------------Ansari-Bardley Test-------------------------------**

Let

H0: The true ratio of scale is equal to 1

H1 :The true ratio of scale is not equal to 1



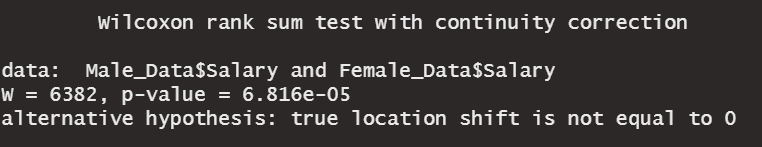
p-value = 0.0009319 <<0.05, we reject null hypothesis at the 5% significance level that the variance in salary is not the same for the two gender.

It did not pass the Ansari-Bradley test, we shall not use the parametric tests method unless we have enough sample sizes. Instead, we use non- parametric tests to compare two means, such as Mann-Whitney-Wilcoxon Test

**------------ Wilcoxon rank sum test with continuity correction-------------**

H0: The median of the population of differences between the paired data is zero.

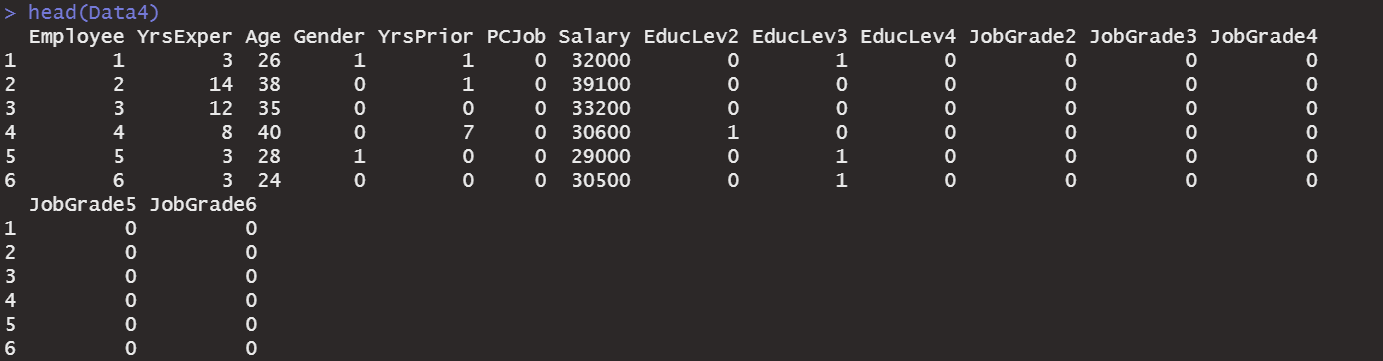
H1: The median of the population of differences between the paired data is not zero.



As alpha = 0.05, p- value <0.05, hence, we reject the null hypothesis. Then we can declare that the average salary of male(median = 42500, IQR = 14250) is not equal to the average salary of female(median= 35450,IQR = 8975). It can support our plaintiff’s claims.

PS: The t-test function was defaulted as Welch, which will consider the variance of this two sample are not the same as default. Hence, the ansari test above is quite redundant anyway.

**Q3:** Transform **EducLev** into several dummy variables. The number of dummy variables you create will need to depend on your logical judgement. Also transform **JobGrade**, **Gender**, and **PCJob** into dummy variables.

****

**Baseline:**

**PCJob: “No” as 0**

Reason: 0 can also represents null.

**Gender: “Female” as 0**

Reason: habitual setting.

**EducLev: EducLev1 as baseline (all 0)**

Reason: set the lowest Education as benchmark for other Education level.

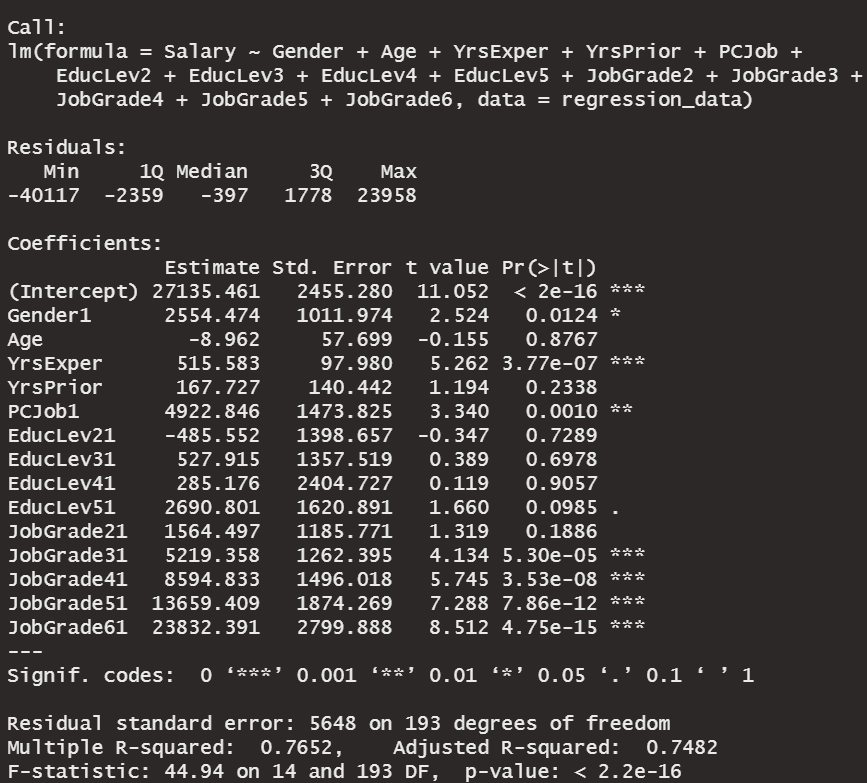
**JobGrade: JobGrade1 as baseline (all 0)**

Reason: set the lowest JobGrade as benchmark for other JobGrade level.

**Q4.** The defense counsel tries to counter against the plaintiff’s argument by showing that the mean difference between the two groups is biased because he or she did not control for several other factors/variables. **Estimate a multiple regression model to strengthen/bolster the plaintiff’s justification,** then **write a report explaining your results**. - Also discuss about: what **R-squared** is and what it means, what the meaning of the **t-values and the coefficients are (or estimates)**.

1. **Estimate a multiple regression model to strengthen/bolster the plaintiff’s justification, then write a report explaining your results.**
2. **(2) What R-squared is and what it means, what the meaning of the t- values and the coefficients are (or estimates). (Below)**

Ans:

****

**Report:**

We use the data4 because it has all dummy variables. I would like to put all the variable into the model first to see the performance.

**Residuals:**

We can take a look for symmetrical distribution across these five statistical points to observe how well our model fit the data. In this case, it may seem like not strongly symmetrical, therefore, it represents that the model predicts certain points that fall far away from the actual observed points.

**coefficients**

**= 27135.461 +**

**Gender1 \* 2554.474+**

**Age \* -8.962+**

**YrsExper \* 515.583+**

**YrsPrior \* 167.727+**

**PCJob1 \* 4922.846+**

**EducLev21 \* -485.552+**

**EducLev31 \* 527.915+**

**EducLev41 \* 285.176+**

**EducLev51 \* 2690.801+**

**JobGrade21 \* 1564.497+**

**JobGrade31 \* 5219.358+**

**JobGrade41 \* 8594.833+**

**JobGrade51 \*13659.409+**

**JobGrade61\* 23832.391**

**R-squared:**

**Range [0,1]**

**In our case, R square = 0.7652, it means that we have approximate 76.52% of the variance x found in the response y can be explained by the predictor variable. (**該變數代表應變數y受自變數x之影響的解釋例程度大小 ,我們可以說該值越接近1，對y變數影響越強)

Then the rest (1- **76.52%**) is the part the model can’t explain.

And for the part of adjusted R square, it’s a special form of the R square that adding additional ‘useful’ data point into model will increase the value, otherwise, it decreases.

**coefficients - Std. Error**

Mostly, we desire for the lower std error. It measures the average amount that the coefficient estimates vary from the actual average value of our variable.

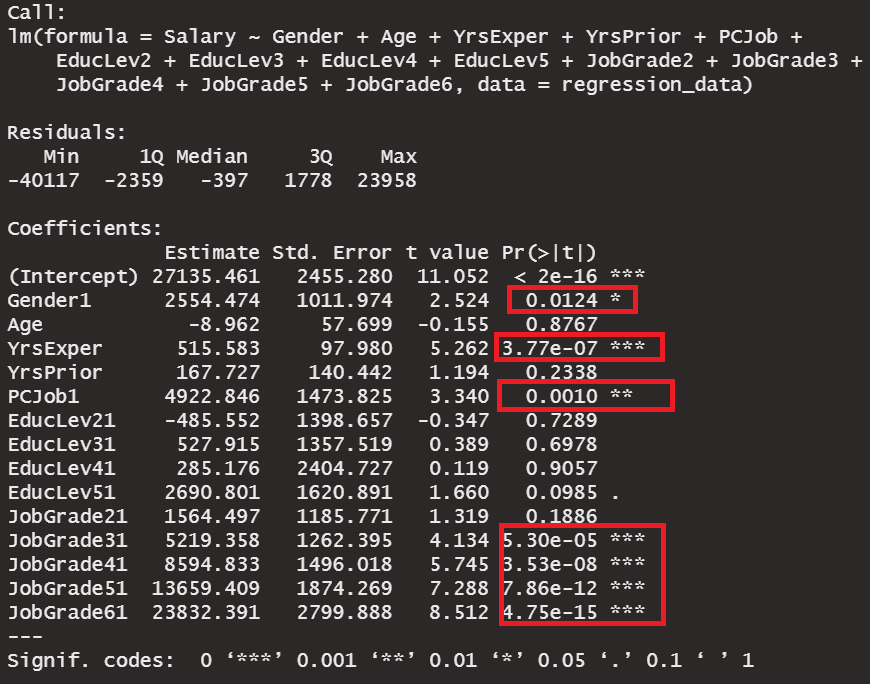
**coefficients – T -value**

It represents that how many standard deviations our estimate value is away from 0. Each T value have its corresponding p-value.

**coefficients – Pr(>|t|)**

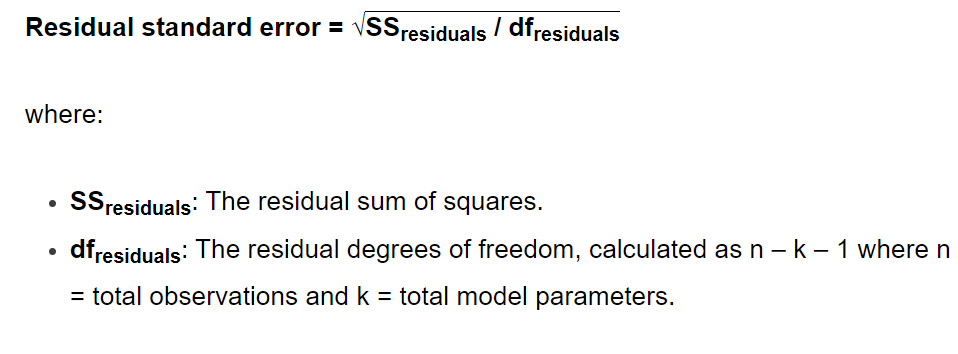
In most cases, we choose α = 0.05 as the benchmark.

In our example, the number in the red box below is less than 0.05.



**Residual Standard Error**

In our case: Residual standard error: 5648 on 193 degrees of freedom. The value of freedom 193 comes from the sample size (n) – model parameters – 1 = 208 -14 -1 = 193



RSE is the average amount that the predictive point vary from the regression line.

**5) Do these data provide evidence that there is discrimination against female employees in terms of salary?**

****

Yes. The p- value of the gender is smaller than 0.05. Therefore, we reject Ho, accept H1. Which means it did have variance between 2 gender.

**Extra Point**

**a. (1) Explain what an interaction term is, (2) how we can estimate a regression model with interaction terms and (3)how we could interpret the results.**

Ans(a):

Interaction occurs as an independent variable's effect on the result depends on the effect of another independent variable. Therefore, we add Interaction term as the variables into our regression model.

For example:

If we only selected 2 significant factors, **YrsExper** and **Gender** , as variables of our model, the term (YrsExper \* Gender) would be the interaction term. In other words, Because of the interaction, the effect of having how much salary for the worker with different past working experience( Year) is different if that employee is male or female.

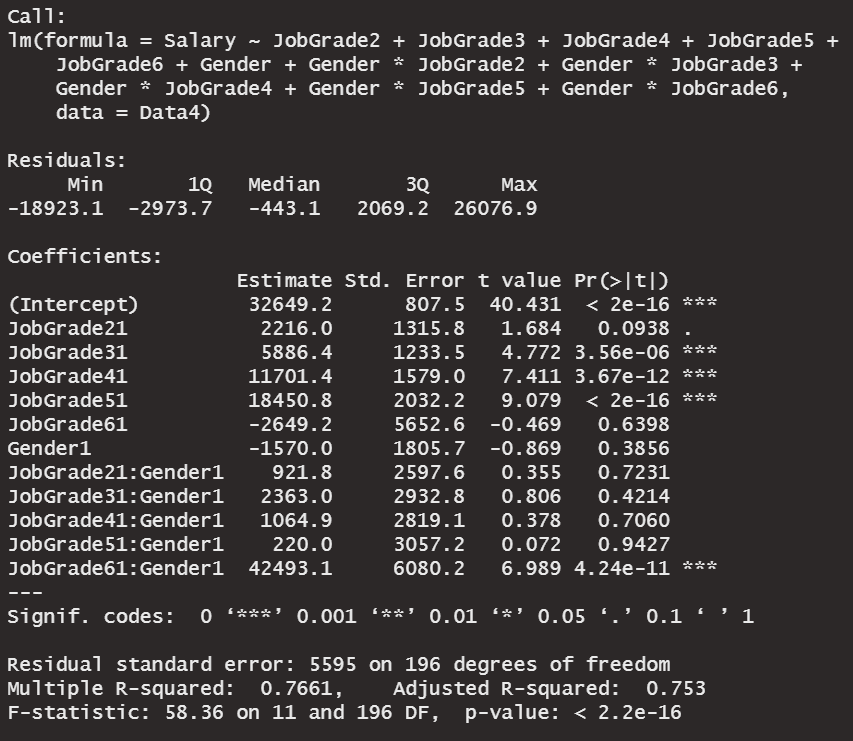
**Salary = intercept + a× YrsExper + b × Gender + c (YrsExper \* Gender) + residual error**

Ans (2) (3):

In order to focus on the impact of Job Grade rank on male workers,

we only remain the Job Grade(2-6) and Gender, which are supposed to increase the explanatory power because we have shrunk the amount of the factors. Furthermore, I want to add the interaction term into the model to observe the male worker(male = 1) in different JobGrade.#Normally, it should be the male with higher rank having the higher salary.

#Using Data4 with dummy variable

****

Because they are 2 category variables, the multiplication of this 2 variables would only remain the observation both represented as 1,

eg, JobGrade1 \*Gender -> male with JobGrade1

eg, JobGrade2 \*Gender -> male with JobGrade2

**b. How would you determine whether the interaction terms contribute in a meaningful way to the explanatory power of your estimation model?**

The explanatory power of the model is depending on the change of the R square. However, the R square could simply boost if we put plenty interaction terms inside the model, which also makes us hard to explain by words how each factor effects the predictive target separately.

In my case, most of my interaction terms are not significant and they only boost my R square from 0.7652 to 0.7661, which is quite low. Therefore, I think that it’s not a meaningful movement.