**Answers to AS3 – PBA**

Daniele Melotti (ID: 110077432)

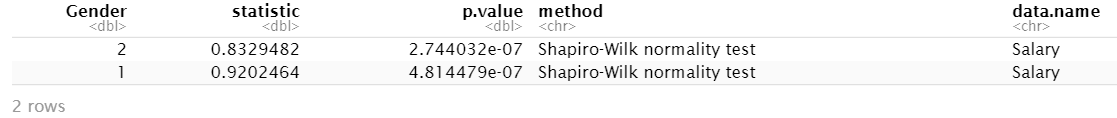
1. First, I converted the Salary variable into numeric after removing the dollar sign and comma, then, I prepared a summary of the variables using summary(bs), where bs is where the bank salary file is stored:

|  |  |  |
| --- | --- | --- |
| **Employee** | **EducLev** | **JobGrade** |
| Min.: 1.00  1st Qu.: 52.75  Median: 104.50  Mean: 104.50  3rd Qu.: 156.25  Max.: 208.00 | Min.: 1.000  1st Qu.: 2.000  Median: 3.000  Mean: 3.159  3rd Qu.: 5.000  Max.: 5.000 | Min.: 1.00  1st Qu.: 1.00  Median: 3.00  Mean: 2.76  3rd Qu.: 4.00  Max.: 6.00 |
| **YrsExper** | **Age** | **Gender** |
| Min.: 2.000  1st Qu.: 5.000  Median: 8.000  Mean: 9.673  3rd Qu.: 13.000  Max.: 39.000 | Min.: 22.00  1st Qu.: 32.00  Median: 38.50  Mean: 40.39  3rd Qu.: 47.25  Max.: 65.00 | Length: 208  Class: character  Mode: character |
| **YrsPrior** | **PCJob** | **Salary** |
| Min.: 0.000  1st Qu.: 0.000  Median: 1.000  Mean: 2.375  3rd Qu.: 4.000  Max.: 18.000 | Length: 208  Class: character  Mode: character | Min.: 26700  1st Qu.: 33000  Median: 37000  Mean: 39922  3rd Qu.: 44000  Max.: 97000 |

1. The plaintiff’s claim would lead to using a two-sample t-test, with the following hypotheses:

vs.

But before doing that, I checked the normality of the distribution of the Gender table using the Shapiro test:



The p-values are both extremely close to zero, meaning that we can reject the null hypothesis and believe that the variable’s distribution is not normal. This means that the Ansari-Bradley test is the most indicated one for testing the equality of variances:

Text

Description automatically generated

The test’s result leads to rejecting the null hypothesis, meaning that the variances are likely to be not equal. Therefore, we can run the t.test, but we must set var.equal to FALSE (it is FALSE by default actually), which will involve using the Welch approximation. The code is the following:

t.test(Salary Gender, data = bs, var.equal = FALSE)

And the output:

Text, letter

Description automatically generated

Considering the very low p-value, we can reject the null hypothesis and confirm the lawyer’s claims. This means that there is a significant difference in average salary between males and females.

1. In order to transform the variables of interest into several dummy variables I used the package fastDummies and the function dummy\_cols:

bs <- dummy\_cols(bs, select\_columns = c("EducLev", "JobGrade", "Gender", "PCJob"), remove\_first\_dummy = TRUE)

Setting remove\_first\_dummy = TRUE helped creating k-1 dummies for each variable. The created dummies are the following: EducLev\_2, EducLev\_3, EducLev\_4, EducLev\_5, JobGrade\_2, JobGrade\_3, JobGrade\_4, JobGrade\_5, JobGrade\_6, Gender\_Male, PCJob\_Yes.

1. I ran the multiple regression using the following code:

reg <- lm(Salary ~ YrsExper + Age + YrsPrior + EducLev\_2 + EducLev\_3 + EducLev\_4 + EducLev\_5 + JobGrade\_2 + JobGrade\_3 + JobGrade\_4 + JobGrade\_5 + JobGrade\_6 + Gender\_Male + PCJob\_Yes, data = bs)

Then I used the stargazer package in order to get the following report as ouput:

A picture containing table

Description automatically generated

The model shows that on average males earn 2,554.47 more than female, other conditions being equal. This difference is statistically significant within a 95% confidence level. In addition to that, we see that JobGrade and PCJob are also very influential on salary. In fact, we see that the higher the job grade, the higher the salary, with people within the 6th grade earning an average extra 23,832.39 over people who possess level 1. Having a computer related job grants an average extra 4,922.846 on people who don’t have a computer-related job, ceteris paribus.

The R-squared is intended as goodness of fit and represents what percentage of the outcome variable is explained by the model. In this model R-squared is 0.765 (76.5%), which is quite good considering that it can range between 0 and 1.

The t values can be calculated dividing the estimated coefficient by its standard error, which is the standard deviation of residuals (the error in estimation). When a coefficient is relatively big compared to its standard error, the t value will also be big, which means that the estimate will probably be statistically significantly different from 0. The t value for Gender\_Male is:

The coefficients indicate the change in the outcome variable for a 1-unit increase of an explanatory variable. For instance, in our model a unitary increase in terms of years of experience (YrsExper) leads to an average increase in salary of 515.58.

1. As mentioned in the answer above, males earn 2,554.47 more than female on average, other conditions being equal. So, there is definitely evidence that there is a discrimination against female employees in terms of salary.

**Extra credit**

1. An interaction is when an input variable has a different effect on the outcome variable depending on the values of another independent variable.

In order to better explain what an interaction term is, I will take a simplified model from the variables used in this assignment. Let’s imagine the following model:

Assume that y is the Salary, Gender is a variable that takes value 0 for males and 1 for females, and that Age indicates the age of the employee. Now, if we are dealing with a male employee, the estimated salary will be equal to:

However, if we are dealing with a female employee, the predicted salary would be:

Let’s set some imaginary value for the variables, such as , and .

Then, substituting in the equation for males, we obtain:

While in the equation for females we obtain:

Surely, the equations for males and females have different intercept after substituting the parameters, but they have the same slope. If we drew the plots, we would see two parallel lines

This could be seen as a limitation, because it would mean that the salary increases equally for males and females as they age (of course, this would be fair, but given the plaintiff’s case it might not be the reality of things). For some reason, maybe females earn more than males as they age. The current model cannot say this. But the introduction of an interaction term can help. The starting model would become:

Let’s assume that . Now substituting the data into the equations will be the best way to visualize the effect of the interaction term. For males:

We see that the outcome for males is the same as before. But for females:

The equations have now a different slope, showing that females would earn more compared to males when they age!

I tried running a regression with the interaction term between Gender\_Male and Age:

inter <- lm(Salary ~ YrsExper + Age + YrsPrior + EducLev\_2 + EducLev\_3 + EducLev\_4 + EducLev\_5 + JobGrade\_2 + JobGrade\_3 + JobGrade\_4 + JobGrade\_5 + JobGrade\_6 + Gender\_Male + PCJob\_Yes + Gender\_Male \* Age, data = bs)

The output:

A picture containing table

Description automatically generated

As we can see, the level of has increased a little. The interaction term is statistically significant with 99% confidence, however, the coefficient for Gender\_Male has become strongly negative, meaning that males earn 8,710.49 – 298.985 = 8411.504 less than females on average, ceteris paribus.

1. I would see if the of the model with the interaction term has increased compared to the one of the model with no interaction term. Also, I would check that the estimate for the interaction term is statistically significant.