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

2) The plaintiff's claim would lead to using a two-sample t-test, with the following hypotheses:

$$H_0: \mu_{salary}^{male} = \mu_{salary}^{female} \text{ vs. } H_a: \mu_{salary}^{male} \neq \mu_{salary}^{female}$$

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

Gender <dbl>	statistic <dbl>	p.value <dbl>	method <chr>	data.name <chr>
2	0.8329482	2.744032e-07	Shapiro-Wilk normality test	Salary
1	0.9202464	4.814479e-07	Shapiro-Wilk normality test	Salary

2 rows

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:

```
        Ansari-Bradley test

data:  Salary by Gender
AB = 8024, p-value = 0.0009319
alternative hypothesis: true ratio of scales is not equal to 1
```

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:

```
        Welch Two Sample t-test

data:  Salary by Gender
t = -4.141, df = 78.898, p-value = 8.604e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12282.943  -4308.082
sample estimates:
mean in group Female   mean in group Male
            37209.93             45505.44
```

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.

3) 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.

4) 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:

	Dependent variable:
	Salary
YrsExper	515.583***
	(97.980)
Age	-8.962
	(57.699)
YrsPrior	167.727
	(140.442)
EducLev_2	-485.552
	(1,398.657)
EducLev_3	527.915
	(1,357.519)
EducLev_4	285.176
	(2,404.727)
EducLev_5	2,690.801*
	(1,620.891)
JobGrade_2	1,564.497
	(1,185.771)
JobGrade_3	5,219.358***
	(1,262.395)
JobGrade_4	8,594.833***
	(1,496.018)
JobGrade_5	13,659.410***
	(1,874.269)
JobGrade_6	23,832.390***
	(2,799.888)
Gender_Male	2,554.474**
	(1,011.974)
PCJob_Yes	4,922.846***
	(1,473.825)
Constant	27,135.460***
	(2,455.280)
Observations	208
R2	0.765
Adjusted R2	0.748
Residual Std. Error	5,648.080 (df = 193)
F Statistic	44.939*** (df = 14; 193)
Note:	*p<0.1; **p<0.05; ***p<0.01

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 6^{th} 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:

$$t_{value} = \frac{2{,}554.474}{1{,}011.974} = 2.52$$

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.

5) 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

a) 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:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$y = \beta_0 + \beta_{gender} x_{gender} + \beta_{Age} x_{Age}$$

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:

$$y_{male} = \beta_0 + \beta_{gender} \cdot 0 + \beta_{Age} x_{age}$$

$$y = \beta_0 + \beta_{Age} x_{age}$$

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

$$y_{female} = \beta_0 + \beta_{gender} x_{gender} + \beta_{Age} x_{age}$$

Let's set some imaginary value for the variables, such as $\beta_0 = 100$, $\beta_{gender} = 10$ and $\beta_{age} = 2$. Then, substituting in the equation for males, we obtain:

$$y_{male} = 100 + 10 \cdot 0 + 2 \cdot x_{age}$$

$$y_{male} = 100 + 2 \cdot x_{age}$$

While in the equation for females we obtain:

$$y_{female} = 100 + 10 \cdot 1 + 2 \cdot x_{age}$$

$$y_{female} = 110 + 2 \cdot x_{age}$$

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:

$$y = \beta_0 + \beta_{gender} x_{gender} + \beta_{Age} x_{Age} + \beta_{interaction} x_{gender} x_{age}$$

Let's assume that $\beta_{interaction} = 1$. Now substituting the data into the equations will be the best way to visualize the effect of the interaction term. For males:

$$y_{male} = \beta_0 + \beta_{gender} \cdot 0 + \beta_{Age} x_{age} + \beta_{interaction} \cdot 0 \cdot x_{age}$$

$$y_{male} = \beta_0 + \beta_{Age} x_{age}$$

$$y_{male} = 100 + 2 \cdot x_{age}$$

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

$$y_{female} = \beta_0 + \beta_{gender} x_{gender} + \beta_{Age} x_{age} + \beta_{interaction} x_{gender} x_{age}$$

$$y_{female} = 100 + 10 \cdot 1 + 2 \cdot x_{age} + 1 \cdot 1 \cdot x_{age}$$

$$y_{female} = 110 + 2 \cdot x_{age} + 1 \cdot x_{age}$$

$$y_{female} = 110 + 3 \cdot x_{age}$$

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:

	Dependent variable:
	Salary
YrsExper	506.000***
	(95.286)
Age	-108.835*
	(62.937)
YrsPrior	90.005
	(138.320)
EducLev_2	-209.890
	(1,361.918)
EducLev_3	308.225
	(1,321.138)
EducLev_4	206.791
	(2,337.744)
EducLev_5	2,516.268
	(1,576.460)
JobGrade_2	1,823.770
	(1,155.070)
JobGrade_3	5,435.729***
	(1,228.733)
JobGrade_4	8,923.928***
	(1,457.320)
JobGrade_5	13,511.530***
	(1,822.469)
JobGrade_6	20,643.410***
	(2,870.382)
Gender_Male	-8,710.489**
	(3,367.160)
PCJob_Yes	4,815.672***
	(1,433.034)
Age:Gender_Male	298.985***
	(85.469)
Constant	31,456.900***
	(2,687.525)
Observations	208
R2	0.779
Adjusted R2	0.762
Residual Std. Error	5,490.503 (df = 192)
F Statistic	45.201*** (df = 15; 192)
Note:	*p<0.1; **p<0.05; ***p<0.01

As we can see, the level of R^2 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.

b) I would see if the R^2 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.