|  |
| --- |
| SFU |
|  |
| BUEC 333 Assignment 2 |
| Kun Yang 301178299 |
|  |
|  |
|  |

|  |
| --- |
|  |

I attached the R codes and the copy of outputs of R in each question. And I am a STAT minor student, so I use the knowledge of R from my previous STAT courses (STAT 445 and 475). I also attached the outputs of the pre-requirements of this assignment in Appendix A. (“summary”, “str”, “head")

1. ***Before we start the assignment, enter the following two lines of code and copy the output that R returns into your assignment.***

> timestamp()

##------ Mon Jul 21 21:26:55 2014 ------##

> getwd()

[1] "C:/Users/Kun/Documents"

1. ***There is no variable that contains the wage, as only the log of wage is provided. Generate a new variable, “unionData$wage", that contains the hourly wage in US dollars.***

> unionData$wage<-exp(unionData$lwage)

> unionData$wage

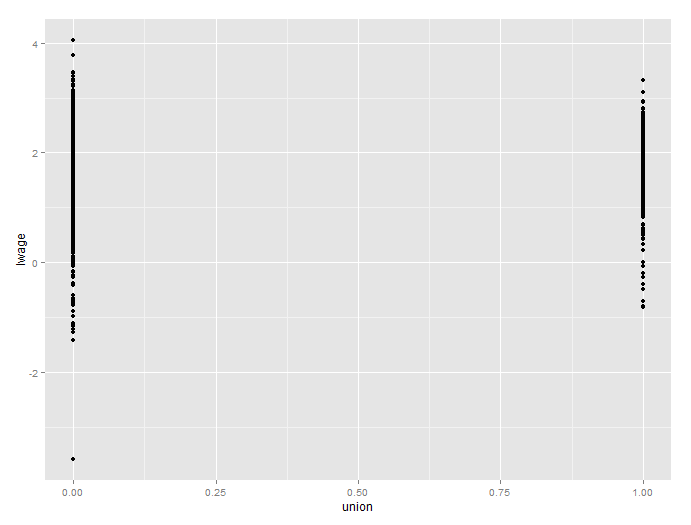
## I do not think this question need to be explained so I attached parts of the outputs in appendix B. You can check it.

1. ***Make a scatterplot (you can use the package “ggplot", for example) of “wage" and the variable “union", which measures whether somebody is a part of a union or not. What do you conclude from this scatterplot? Does it make a difference when you do it for the log wage?***

> require(ggplot2)

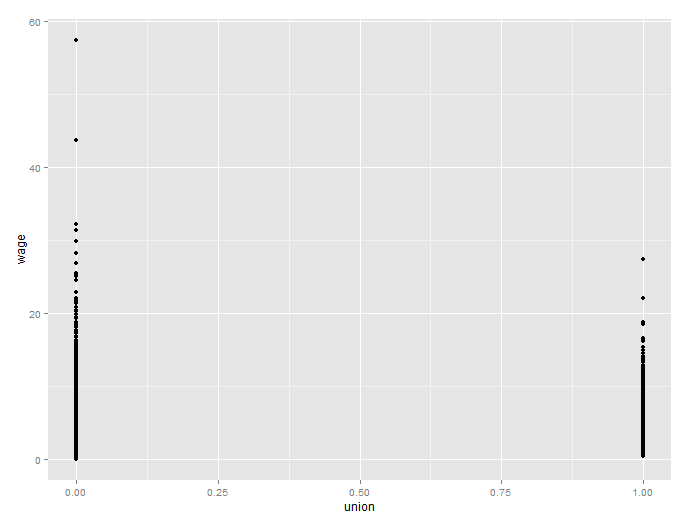
*# scatterplot of “log wage" and “union"*

> qplot(union,lwage,data=unionData)



*# scatterplot of “wage" and “union"*

> qplot(union,wage,data=unionData)



**Answers in 2-3 lines:**

In the “wage” and “union” plot, the non-union wages skewed to have higher points and there are also two extremely high wages in non-union group. But in “log wage” and “union” plot, the non-union wages are not skewed but more spread out than wages bargained by union.

1. ***Run a regression of “wage" on “union". Interpret the regression coefficient estimate for “union". Report the 90% confidence interval for the regression coefficient estimate of union: what do you conclude?***

> fit1=lm(wage~union,data=unionData)

> summary(fit1)

Call:

lm(formula = wage ~ union, data = unionData)

Residuals:

Min 1Q Median 3Q Max

-6.132 -2.013 -0.598 1.345 51.797

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.70771 0.05541 103.017 < 2e-16 \*\*\*

union 0.86652 0.11216 7.726 1.37e-14 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.181 on 4358 degrees of freedom

Multiple R-squared: 0.01351, Adjusted R-squared: 0.01329

F-statistic: 59.69 on 1 and 4358 DF, p-value: 1.368e-14

> confint(fit1,'union',level=0.90)

5 % 95 %

union 0.6820009 1.051043

**Answers in 2-3 lines:**

The wage is estimated 0.87 higher for the employees who are part of union than who are not, ceteris paribus. This confidence interval does not contain 0, so the effect of “union” on “wage” is significant at confidence level of 90%.

1. ***Do the same for “lwage".***

> fit2=lm(lwage~union,data=unionData)

> summary(fit2)

Call:

lm(formula = lwage ~ union, data = unionData)

Residuals:

Min 1Q Median 3Q Max

-5.1845 -0.2903 0.0197 0.3321 2.4465

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.605400 0.009181 174.866 <2e-16 \*\*\*

union 0.179264 0.018584 9.646 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5271 on 4358 degrees of freedom

Multiple R-squared: 0.0209, Adjusted R-squared: 0.02068

F-statistic: 93.04 on 1 and 4358 DF, p-value: < 2.2e-16

> confint(fit2,'union',level=0.90)

5 % 95 %

union 0.148689 0.2098394

**Answers in 2-3 lines:**

For the employees who are part of union the log wage is estimated 0.17 higher than those who are not in union, ceteris paribus. This confidence interval does not contain 0, so the effect of “union” on “wage” is significant at confidence level of 90%

1. ***I prefer the model with “lwage" over the model with “wage". However, there is a problem because we did not include any variables other than “union". Explain what the problems are with this, from a theoretical/statistical point of view.***

**Answers in 2-3 lines:**

There may exist omitted variable bias because there are some other variables like experience, education, etc. Those variables are determinants of wage and may correlate with “union”. Thus, the “union” effect may also contain the effect of other variables.

1. ***Run a regression of “lwage" on union, hours, year, occ1, occ2, occ3, occ4, occ5, occ6, occ7,occ8, occ9. R refuses to give an estimate for the regression coefficient on occ9. Why?***

> fit3=lm(lwage~union+hours+year+occ1+occ2+occ3+occ4+occ5+occ6+occ7+occ8+occ9,data=unionData)

> summary(fit3)

Call:

lm(formula = lwage ~ union + hours + year + occ1 + occ2 + occ3 + occ4 + occ5 + occ6 + occ7 + occ8 + occ9, data = unionData)

Residuals:

Min 1Q Median 3Q Max

-5.4186 -0.2533 0.0293 0.3059 2.1956

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.166e+02 6.674e+00 -17.470 < 2e-16 \*\*\*

union 2.295e-01 1.780e-02 12.890 < 2e-16 \*\*\*

hours -5.909e-05 1.379e-05 -4.286 1.86e-05 \*\*\*

year 5.954e-02 3.368e-03 17.677 < 2e-16 \*\*\*

occ1 4.226e-01 3.199e-02 13.212 < 2e-16 \*\*\*

occ2 3.740e-01 3.363e-02 11.121 < 2e-16 \*\*\*

occ3 3.183e-01 3.926e-02 8.107 6.66e-16 \*\*\*

occ4 1.887e-01 3.116e-02 6.057 1.50e-09 \*\*\*

occ5 2.699e-01 2.727e-02 9.897 < 2e-16 \*\*\*

occ6 1.828e-01 2.752e-02 6.643 3.45e-11 \*\*\*

occ7 8.913e-02 3.284e-02 2.714 0.00666 \*\*

occ8 -4.992e-02 6.588e-02 -0.758 0.44865

occ9 NA NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.491 on 4348 degrees of freedom

Multiple R-squared: 0.1524, Adjusted R-squared: 0.1502

F-statistic: 71.05 on 11 and 4348 DF, p-value: < 2.2e-16

**Answers in 2-3 lines:**

There is no coefficient on occ9 because of singularities. It means that the variable occ9 is depends on occ1-8. And if occ1-8 all equal to 0 then occ9 will be 1. Therefore, it is meaningless to contain occ9 in this regression model and it was automatically ignored by R.

1. ***Interpret the estimate of the regression coefficient of “year".***

**Answers in 2-3 lines:**

The interpretation of the estimate of the regression coefficient of “year” is that we can estimate that taking one additional year of teaching decreases log wage hourly by 5.909e-05, holding other variables constant.

1. ***I wonder whether I should include a person's work experience, as measured by the variable “exper". Try it, and argue why it should or should not be included. You can be informal.***

### occ9 can be eliminated without changing the result and occ8 also seems invalid based on the outputs above. But the question did not ask me to delete occ8, thus I just keep it there. The model shows as bellowing.

> fit4=lm(lwage~union+hours+year+occ1+occ2+occ3+occ4+occ5+occ6+occ7+occ8+exper,data=unionData)

> summary(fit4)

Call:

lm(formula = lwage ~ union + hours + year + occ1 + occ2 + occ3 + occ4 + occ5 + occ6 + occ7 + occ8 + exper, data = unionData)

Residuals:

Min 1Q Median 3Q Max

-5.4205 -0.2571 0.0359 0.3041 2.1743

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.452e+02 1.147e+01 -12.654 < 2e-16 \*\*\*

union 2.294e-01 1.779e-02 12.897 < 2e-16 \*\*\*

hours -5.777e-05 1.378e-05 -4.192 2.82e-05 \*\*\*

year 7.399e-02 5.798e-03 12.762 < 2e-16 \*\*\*

occ1 4.106e-01 3.219e-02 12.753 < 2e-16 \*\*\*

occ2 3.670e-01 3.368e-02 10.898 < 2e-16 \*\*\*

occ3 3.126e-01 3.927e-02 7.960 2.18e-15 \*\*\*

occ4 1.848e-01 3.116e-02 5.933 3.21e-09 \*\*\*

occ5 2.762e-01 2.733e-02 10.108 < 2e-16 \*\*\*

occ6 1.881e-01 2.755e-02 6.827 9.84e-12 \*\*\*

occ7 9.454e-02 3.285e-02 2.878 0.00403 \*\*

occ8 -3.831e-02 6.593e-02 -0.581 0.56125

exper -1.430e-02 4.672e-03 -3.061 0.00222 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4905 on 4347 degrees of freedom

Multiple R-squared: 0.1542, Adjusted R-squared: 0.1518

F-statistic: 66.03 on 12 and 4347 DF, p-value: < 2.2e-16

**Answers in 2-3 lines:**

“exper” needs to be included because it is the determinate of wage and seems affect other variables. Based on t-test, the null hypothesis was rejected and “exper” has effect on wage. the adjusted R^2 increased from 0.1502 to 0.1518. However, the negative sign is suspicious.

1. ***The relationship between wages and experience seems suspicious. There may still be an omitted variable problem. Include “educ". What do you conclude about education and experience, and their relationship with wages?***

###The regression model between log wage and other variables shows as following, again I did not delete OCC8 because you did not ask me to do but it seems weird.

> fit5=lm(lwage~union+hours+year+occ1+occ2+occ3+occ4+occ5+occ6+occ7+occ8+exper+educ,data=unionData)

> summary(fit5)

Call:

lm(formula = lwage ~ union + hours + year + occ1 + occ2 + occ3 +

occ4 + occ5 + occ6 + occ7 + occ8 + exper + educ, data = unionData)

Residuals:

Min 1Q Median 3Q Max

-5.4042 -0.2364 0.0384 0.2948 2.3645

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.697e+01 1.238e+01 -4.601 4.32e-06 \*\*\*

union 2.064e-01 1.733e-02 11.911 < 2e-16 \*\*\*

hours -6.772e-05 1.339e-05 -5.057 4.44e-07 \*\*\*

year 2.888e-02 6.272e-03 4.604 4.26e-06 \*\*\*

occ1 3.052e-01 3.192e-02 9.563 < 2e-16 \*\*\*

occ2 2.938e-01 3.301e-02 8.903 < 2e-16 \*\*\*

occ3 2.419e-01 3.837e-02 6.305 3.16e-10 \*\*\*

occ4 1.561e-01 3.030e-02 5.152 2.69e-07 \*\*\*

occ5 2.775e-01 2.653e-02 10.459 < 2e-16 \*\*\*

occ6 1.975e-01 2.675e-02 7.384 1.84e-13 \*\*\*

occ7 1.070e-01 3.190e-02 3.352 0.000808 \*\*\*

occ8 -2.834e-02 6.401e-02 -0.443 0.657980

exper 3.317e-02 5.390e-03 6.154 8.21e-10 \*\*\*

educ 8.691e-02 5.331e-03 16.302 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4762 on 4346 degrees of freedom

Multiple R-squared: 0.2029, Adjusted R-squared: 0.2005

F-statistic: 85.11 on 13 and 4346 DF, p-value: < 2.2e-16

**Answers in 2-3 lines:**

Experience and education affect each other and the adjusted R^2 increased from 0.15 to 0.20. For t-test, both null hypotheses were rejected. So we add both variables. As the year of experience and education increase by 1, the estimated log wage, respectively, increase by 0.03 and 0.08, holding others constant.

1. ***Are unions good for employees' wages?***

**Answers in 2-3 lines:**

Yes, I would like to say that unions are good for wages. Based on the model showed above, the log wages are estimated 0.2 higher for those who join the unions compared to those who are not part of the unions, holding others constant.