# **W271**-2 – Spring 2016 – **HW 3**

**Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song**

February 17, 2016

## Contents



---

## Exercises

**Complete the following exercises, following the best practices outlined in class. Place your answers in a written report (pdf, word, or jupyter notebook format) along with relevant R statements and output.**

### Question 1

**Load the twoyear.RData dataset and describe the basic structure of the data.**

```
##    variable                          label
## 1    female                    =1 if female
## 2   phsrank  % high school rank; 100 = best
## 3        BA         =1 if Bachelor's degree
## 4        AA        =1 if Associate's degree
## 5     black          =1 if African-American
## 6  hispanic                  =1 if Hispanic
## 7        id                      ID Number
## 8     exper  total (actual) work experience
## 9        jc            total 2-year credits
## 10     univ            total 4-year credits
## 11    lwage                 log hourly wage
## 12   stotal   total standardized test score
## 13   smcity          =1 if small city, 1972
## 14  medcity           =1 if med. city, 1972
```

```
## 15   submed    =1 if suburb med. city, 1972
## 16   lgcity          =1 if large city, 1972
## 17    sublg   =1 if suburb large city, 1972
## 18  vlgcity     =1 if very large city, 1972
## 19   subvlg =1 if sub. very lge. city, 1972
## 20       ne                 =1 if northeast
## 21       nc             =1 if north central
## 22    south                      =1 if south
## 23  totcoll                        jc + univ
```

```
summary(data)
```

```
##      female          phsrank            BA               AA
##  Min.   :0.0000   Min.   : 0.00   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:44.00   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :1.0000   Median :50.00   Median :0.0000   Median :0.00000
##  Mean   :0.5196   Mean   :56.16   Mean   :0.3065   Mean   :0.04406
##  3rd Qu.:1.0000   3rd Qu.:76.00   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :99.00   Max.   :1.0000   Max.   :1.00000
##      black           hispanic             id            exper
##  Min.   :0.00000   Min.   :0.00000   Min.   :   19   Min.   :  3.0
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:19372   1st Qu.:104.0
##  Median :0.00000   Median :0.00000   Median :39301   Median :129.0
##  Mean   :0.09508   Mean   :0.04687   Mean   :40616   Mean   :122.4
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:58842   3rd Qu.:149.0
##  Max.   :1.00000   Max.   :1.00000   Max.   :89958   Max.   :166.0
##        jc              univ           lwage            stotal
##  Min.   :0.0000   Min.   :0.000   Min.   :0.5555   Min.   :-3.32480
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.9253   1st Qu.:-0.32734
##  Median :0.0000   Median :0.200   Median :2.2763   Median : 0.00000
##  Mean   :0.3389   Mean   :1.926   Mean   :2.2481   Mean   : 0.04748
##  3rd Qu.:0.0000   3rd Qu.:4.200   3rd Qu.:2.5969   3rd Qu.: 0.61079
##  Max.   :3.8333   Max.   :7.500   Max.   :3.9120   Max.   : 2.23537
##      smcity          medcity           submed           lgcity
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.2854   Mean   :0.1174   Mean   :0.06861   Mean   :0.09448
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##      sublg            vlgcity            subvlg             ne
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.0000
##  Mean   :0.08709   Mean   :0.05855   Mean   :0.06358   Mean   :0.2107
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##        nc              south           totcoll
##  Min.   :0.0000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.000
##  Median :0.0000   Median :0.0000   Median : 1.507
##  Mean   :0.2988   Mean   :0.3271   Mean   : 2.265
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 4.367
##  Max.   :1.0000   Max.   :1.0000   Max.   :10.067
```

The twoyear dataset contains 6763 observations of 23 variables related to wages, educational attainment, and respondent demographics.

## Question 2

**Typically, you will need to thoroughly analyze each of the variables in the data set using univariate, bivariate, and multivariate analyses before attempting any model. For this homework, assume that this step has been conducted. Estimate the following regression:**

$$\begin{aligned}\mathbf{log(wage)} = \beta_0 + \beta_1\mathbf{jc} + \beta_2\mathbf{univ} + \beta_3\mathbf{exper} + \beta_4\mathbf{black} + \beta_5\mathbf{hispanic} \\ + \beta_6\mathbf{AA} + \beta_7\mathbf{BA} + \beta_8\mathbf{exper}\cdot\mathbf{black} + \mathbf{e}\end{aligned}$$

**Interpret the coefficients $\hat{\beta}_4$ and $\hat{\beta}_8$.**

```
model$coefficients[5]
```

```
##      black
## 0.03317088
```

```
model$coefficients[9]
```

```
##  exper:black
## -0.001267898
```

The expected logged wages of black respondents, holding other variables constant, is $\beta_0 + \beta_4$ or 1.45. The coefficient for $\beta_4$ represents the difference in logged wages for a black respondent versus a non-black respondent. The value of $\beta_4$, 0.03, is not statistically significant ($\beta_4 = 0.03$, t=0.54, p=n.s.).

The expected logged wages of respondents holding a bachelor's degree, holding other variables constant, is $\beta_0 + \beta_8$, or 1.50. The coefficient for $\beta_8$ represents the difference in logged wages for a respondent with a bachelor's degree versus a respondent without a bachelor's degree. The value of $\beta_8$, 0.02, is not statistically significant (beta_8 = 0.02, t=1.13, p=n.s.).

## Question 3

**With this model, test that the return to university education is 7%**

To test that the return to university education is 7%, we set up the following hypotheses:

$$H_0 : \beta_2 = 0.7$$
$$H_A : \beta_2 \neq 0.7$$

To obtain the t-statistic we use the following formula:

$$t = \frac{\hat{\beta} - H_0}{\hat{stderr}}$$

```
##     univ
## 1.041918
```

```
p_value
```

```
##      univ
## 0.2974869
```

The coefficient for university is statistically different from 0.7 (t = -199, df = 6754, p < .001)

## Question 4

**With this model, test that the return to junior college education is equal for black and non-black.**

```
# non-black.
```

Given that this model does not include the interaction term of interest, we can only test if intercept for junior college is the same for black and non-black respondents.

```
## Linear hypothesis test
##
## Hypothesis:
## black = 0
##
## Model 1: restricted model
## Model 2: lwage ~ jc + univ + exper + black + hispanic + AA + BA + exper *
##     black
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1   6755
## 2   6754  1 0.2329 0.6294
```

The differece in intercepts is not significantly diferent (F(1,6754) = 0.23, p = n.s.), suggesting that the returns for junior college for black and non black respondents are not different.

## Question 5

**With this model, test whether the return to univeristy education is equal to the return to 1 year of working experience.**

```
# 1 year of working experience.
```

Using the same approach to question 4 we can test the following hypotheses:

$$H_0 : \beta_2 = 12 * \beta_3$$
$$H_A : \beta_2 \neq 12 * \beta_3$$

```
## Linear hypothesis test
##
## Hypothesis:
## univ - 12 exper = 0
##
## Model 1: restricted model
## Model 2: lwage ~ jc + univ + exper + black + hispanic + AA + BA + exper *
##     black
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   6755
## 2   6754  1 11.968 0.0005445 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Question 6

**Test the overall significance of this regression.**

Looking at the regression output, we can test the hypothesis:

$$H_0 : \beta_i = 0 \forall i$$

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper + black + hispanic + AA +
##     BA + exper * black, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11612 -0.27836  0.00432  0.28676  1.76811
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.4773315  0.0223780  66.017  < 2e-16 ***
## jc           0.0637926  0.0079034   8.072 8.15e-16 ***
## univ         0.0732806  0.0031486  23.274  < 2e-16 ***
```

```
## exper        0.0050234  0.0001667  30.141  < 2e-16 ***
## black        0.0331709  0.0613984   0.540   0.5890
## hispanic    -0.0193629  0.0248914  -0.778   0.4367
## AA          -0.0077759  0.0295497  -0.263   0.7924
## BA           0.0176735  0.0156553   1.129   0.2590
## exper:black -0.0012679  0.0004991  -2.541   0.0111 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4287 on 6754 degrees of freedom
## Multiple R-squared:  0.2282, Adjusted R-squared:  0.2272
## F-statistic: 249.6 on 8 and 6754 DF,  p-value: < 2.2e-16
```

We reject the hypothesis that none of the coefficients in the model is statistically different from zero (f(8, 6754) = 249.6, p < 0.001)

## Question 7

**including a square term of working experience to the regression model built above, estimate the linear regression model again. What is the estimated return to work experience in this model?**

```
# estimate the linear regression model again.
# What is the estimated return to work experience in this model?
```

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper + exper2 + black + hispanic +
##     AA + BA + exper * black, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11982 -0.27743  0.00475  0.28741  1.77397
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.510e+00  4.427e-02  34.108  < 2e-16 ***
## jc           6.417e-02  7.916e-03   8.106 6.14e-16 ***
## univ         7.382e-02  3.211e-03  22.992  < 2e-16 ***
## exper        4.301e-03  8.588e-04   5.008 5.64e-07 ***
## exper2       3.379e-06  3.939e-06   0.858   0.3911
## black        2.994e-02  6.152e-02   0.487   0.6265
## hispanic    -1.932e-02  2.489e-02  -0.776   0.4378
## AA          -7.539e-03  2.955e-02  -0.255   0.7986
## BA           1.797e-02  1.566e-02   1.147   0.2513
## exper:black -1.239e-03  5.002e-04  -2.477   0.0133 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4287 on 6753 degrees of freedom
## Multiple R-squared:  0.2282, Adjusted R-squared:  0.2272
## F-statistic: 221.9 on 9 and 6753 DF,  p-value: < 2.2e-16
```

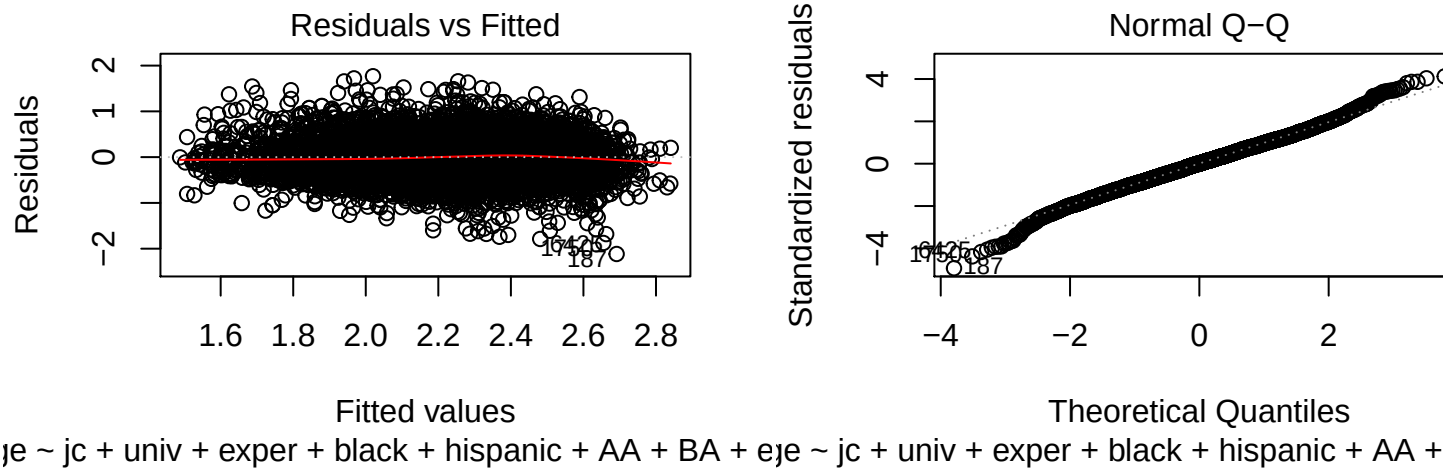
With inclusion of the square of the return to work experience, the coefficient of the return to work experience is 0.004, which is a statistically significant increase in wages compared to workers with no experience (beta = 0.004, t(6753), p <.001). Inclusion of the square term lowers the coefficient for work experience by ~ 0.001.

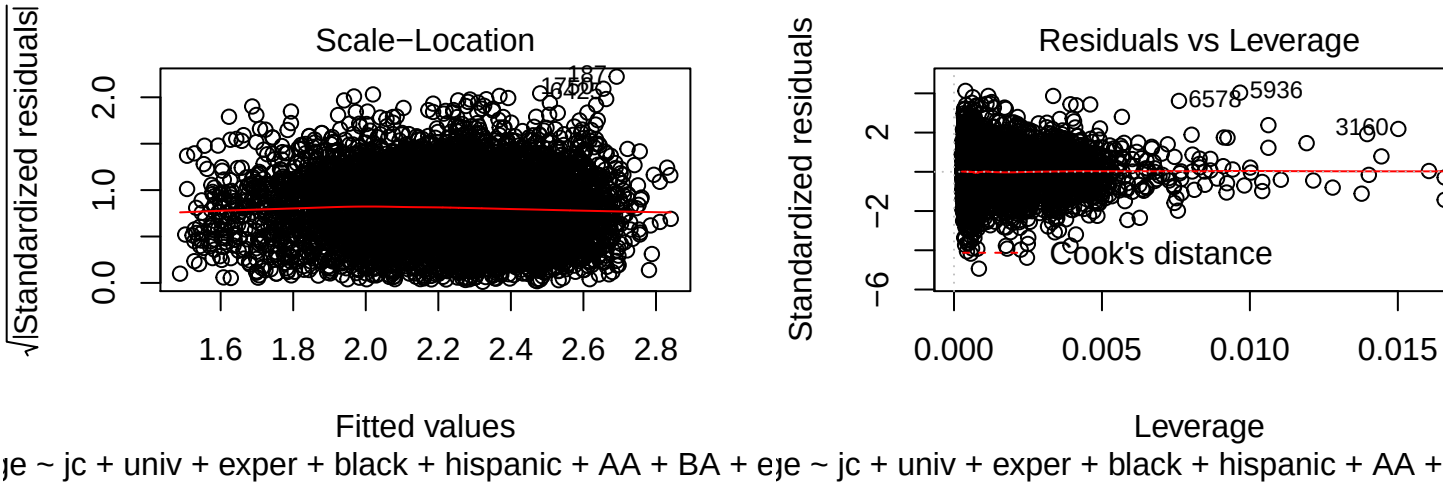
## Question 8

**Provide the diagnosis of the homoskedasticity assumption. Does this assumption hold? If so, how does it affect the testing of no effect of university education on salary change? If not, what potential remedies are available?**

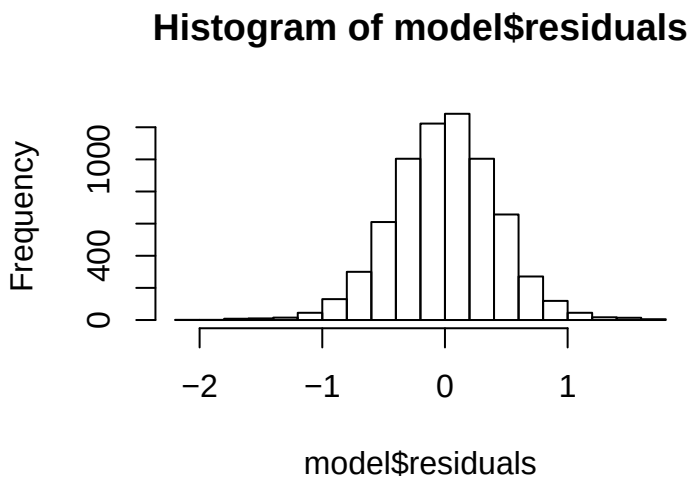
```
# Does this assumption hold?
# If so, how does it affect the testing of no effect of university education on
# salary change?
# If not, what potential remedies are available?
```

```
plot(model)
```



```
hist(model$residuals)
```




Looking at the residuals versus fitted plot, the residuals do not appear to change in distribution at different values of logged wages. The q-q- plot shows that residuals within +- 3 standard deviations generally follow a normal distribution. Similarly, a histogram of the residuals looks generally normal. Given that the sample size is quiet large, normality tests find significant deviation from normallacy, but I would generally conclude the assumption of homoskedasticity is not violated in this case.

In the case of heteroskedasticity, we could no longer assume that our OLS estimates have the smallest possible variance among unbiased, linear estimators and we could not reliably estimate the variance of our coefficients. In this case, the use of robust standard errors would be appropriate.