W271-2 – Spring 2016 – **HW 2**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

February 10, 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.

```
desc
```

```
##    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
```

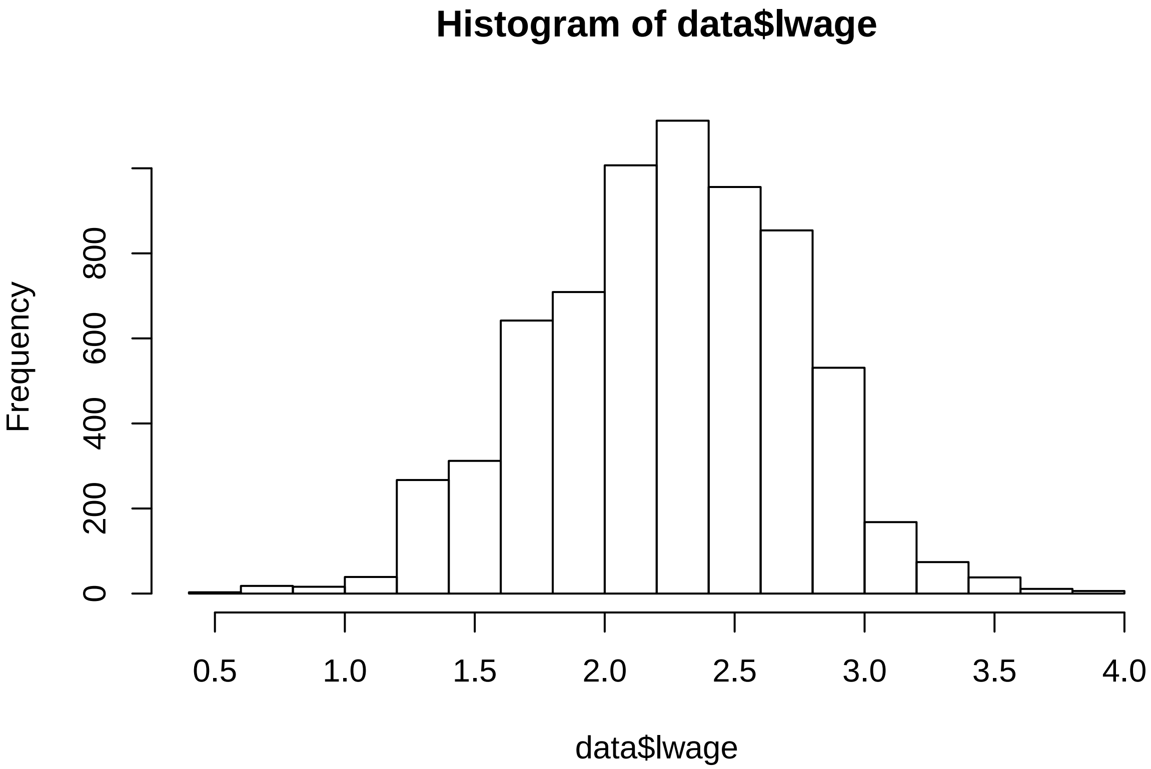
```
summary(data)
```

```
##      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
```

```
#look at the wage variable
summary(data$lwage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5555  1.9250  2.2760  2.2480  2.5970  3.9120
```

```
#look at the histogram to see if there are any potential extremes
hist(data$lwage)
```



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:

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

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

```r
data$experXblack<-data$exper*data$black
model1<-lm(lwage~jc+univ+exper+black+hispanic+AA+BA+experXblack, data=data)
stargazer(model1, type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                lwage
## -----------------------------------------------
## jc                           0.064***
##                              (0.008)
##
## univ                         0.073***
##                              (0.003)
##
## exper                        0.005***
##                             (0.0002)
##
## black                          0.033
##                              (0.061)
##
## hispanic                      -0.019
##                              (0.025)
##
## AA                            -0.008
##                              (0.030)
##
## BA                             0.018
##                              (0.016)
##
## experXblack                  -0.001**
##                             (0.0005)
##
## Constant                     1.477***
##                              (0.022)
##
## -----------------------------------------------
## Observations                   6,763
## R2                             0.228
## Adjusted R2                    0.227
## Residual Std. Error      0.429 (df = 6754)
## F Statistic          249.553*** (df = 8; 6754)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```r
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  1.47733155  0.02293512 64.4135  < 2e-16 ***
## jc           0.06379261  0.00761208  8.3804  < 2e-16 ***
## univ         0.07328063  0.00336598 21.7709  < 2e-16 ***
## exper        0.00502341  0.00016840 29.8294  < 2e-16 ***
## black        0.03317088  0.06872723  0.4826  0.62936
## hispanic    -0.01936289  0.02498704 -0.7749  0.43842
## AA          -0.00777589  0.02746594 -0.2831  0.77710
## BA           0.01767355  0.01656455  1.0670  0.28603
## experXblack -0.00126790  0.00053779 -2.3576  0.01842 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 3

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

```
coeffs<-coefficients(model1)
coeffs[3]
```

```
##       univ
## 0.07328063
```

The coefficient for the univ variable is .073 which equates to approximately a 7% increase in log(wage) for every increment increase in univ education.

Question 4

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

Question 5

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

Question 6

Test the overall significance of this regression.

```
summary(model1)
```

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper + black + hispanic + AA +
##     BA + experXblack, 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
## experXblack -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
```

The overall significance of this regression is high with a F-statistic of 249.6 and an overall p-value <2.2e-16 indicating that overall this model is performing better than random. However, we can likely increase our r squared value by eliminating some of the factors because several factors individually are non significant.

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?

```
data$experSq<-data$exper^2
model2<-lm(lwage~jc+univ+exper+black+hispanic+AA+BA+experXblack+experSq, data=data)
stargazer(model2, type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                lwage
## -----------------------------------------------
## jc                           0.064***
##                               (0.008)
##
## univ                         0.074***
##                               (0.003)
##
## exper                        0.004***
##                               (0.001)
##
## black                          0.030
##                               (0.062)
##
## hispanic                      -0.019
##                               (0.025)
##
## AA                            -0.008
##                               (0.030)
##
## BA                             0.018
##                               (0.016)
##
## experXblack                  -0.001**
##                               (0.001)
##
## experSq                       0.00000
##                              (0.00000)
```

```
##
## Constant                     1.510***
##                               (0.044)
##
## -----------------------------------------------
## Observations                   6,763
## R2                             0.228
## Adjusted R2                    0.227
## Residual Std. Error      0.429 (df = 6753)
## F Statistic          221.898*** (df = 9; 6753)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
coeftest(model2, vcov = vcovHC)
```

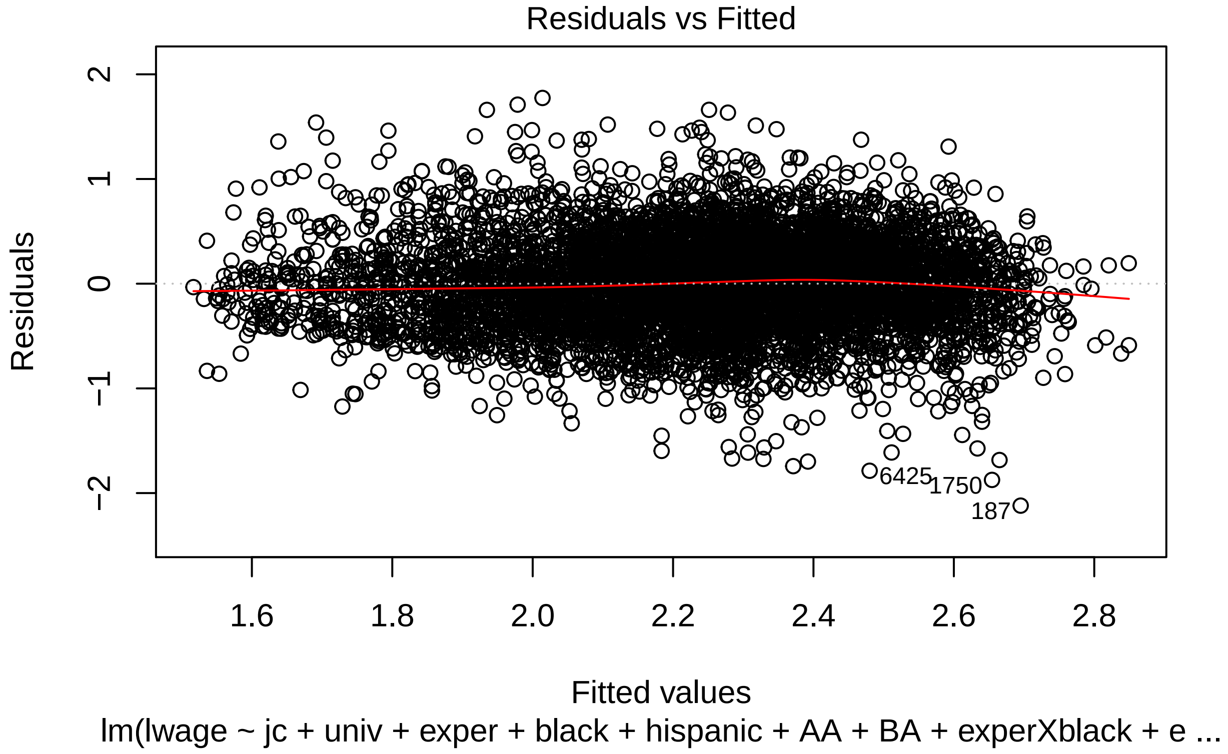
```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  1.5101e+00  4.3591e-02 34.6427 < 2.2e-16 ***
## jc           6.4168e-02  7.6224e-03  8.4183 < 2.2e-16 ***
## univ         7.3819e-02  3.4501e-03 21.3963 < 2.2e-16 ***
## exper        4.3008e-03  8.4541e-04  5.0873 3.731e-07 ***
## black        2.9937e-02  6.8436e-02  0.4374   0.66180
## hispanic    -1.9317e-02  2.4985e-02 -0.7731   0.43947
## AA          -7.5392e-03  2.7481e-02 -0.2743   0.78383
## BA           1.7967e-02  1.6579e-02  1.0837   0.27853
## experXblack -1.2388e-03  5.3539e-04 -2.3139   0.02071 *
## experSq      3.3790e-06  3.8745e-06  0.8721   0.38318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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?

```
plot(model2, which=1)
```

The homoskedasticity assumption may hold in this case. Although the residuals vs fitted plot form the model appears to have a little less variance on the edges (far right and far left) the red line is relatively straight and flat.


Residuals vs Fitted
Residuals
2
1
0
−1
−2
1.6
1.8
2.0
2.2
2.4
2.6
2.8
Fitted values
lm(lwage ~ jc + univ + exper + black + hispanic + AA + BA + experXblack + e ...
6425
1750
187