# **W271**-2 – Spring 2016 – **HW 3**

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

February 17, 2016

## Contents

**Exercises** **2**

Question 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

Question 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

Question 3 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

Question 4 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

Question 5 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7

Question 6 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

Question 7 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

Question 8 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

# Exercises

## Question 1

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

```r
load("twoyear.RData")
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
```

```r
str(data)
```

```
## 'data.frame':    6763 obs. of  23 variables:
##  $ female  : int  1 1 1 1 1 0 0 0 0 0 ...
##  $ phsrank : int  65 97 44 34 80 59 81 50 8 56 ...
##  $ BA      : int  0 0 0 0 0 0 1 0 0 1 ...
##  $ AA      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ black   : int  0 0 0 0 0 0 0 1 0 1 ...
##  $ hispanic: int  0 0 0 1 0 0 0 0 0 0 ...
##  $ id      : num  19 93 96 119 132 156 163 188 199 200 ...
##  $ exper   : int  161 119 81 39 141 165 127 161 138 64 ...
##  $ jc      : num  0 0 0 0.267 0 ...
##  $ univ    : num  0 7.03 0 0 0 ...
##  $ lwage   : num  1.93 2.8 1.63 2.22 1.64 ...
##  $ stotal  : num  -0.442 0 -1.357 -0.19 0 ...
##  $ smcity  : int  0 1 0 1 0 1 1 0 1 0 ...
##  $ medcity : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ submed  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lgcity  : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ sublg   : int  1 0 1 0 0 0 0 0 0 0 ...
##  $ vlgcity : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ subvlg  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ne      : int  1 0 1 0 0 0 0 0 0 0 ...
##  $ nc      : int  0 1 0 0 0 0 1 0 0 0 ...
##  $ south   : int  0 0 0 0 1 1 0 1 0 1 ...
##  $ totcoll : num  0 7.033 0 0.267 0 ...
##  - attr(*, "datalabel")= chr ""
##  - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"
##  - attr(*, "formats")= chr  "%8.0g" "%8.0g" "%8.0g" "%8.0g" ...
##  - attr(*, "types")= int  251 251 251 251 251 251 254 252 254 254 ...
##  - attr(*, "val.labels")= chr  "" "" "" "" ...
##  - attr(*, "var.labels")= chr  "=1 if female" "% high school rank; 100 = best" "=1 if Bachelor's degree" "=1 if
##  - attr(*, "version")= int 10
```

```
head(data)
```

```
##   female phsrank BA AA black hispanic  id exper        jc     univ
## 1      1      65  0  0     0        0  19   161 0.0000000 0.000000
## 2      1      97  0  0     0        0  93   119 0.0000000 7.033333
## 3      1      44  0  0     0        0  96    81 0.0000000 0.000000
## 4      1      34  0  0     0        1 119    39 0.2666667 0.000000
## 5      1      80  0  0     0        0 132   141 0.0000000 0.000000
## 6      0      59  0  0     0        0 156   165 0.0000000 0.000000
##      lwage     stotal smcity medcity submed lgcity sublg vlgcity subvlg ne
## 1 1.925291 -0.4417497      0       0      0      0     1       0      0  1
## 2 2.796494  0.0000000      1       0      0      0     0       0      0  0
## 3 1.625600 -1.3570027      0       0      0      0     1       0      0  1
## 4 2.223312 -0.1900551      1       0      0      0     0       0      0  0
## 5 1.642083  0.0000000      0       0      0      0     0       0      0  0
## 6 2.079442  1.3887565      1       0      0      0     0       0      0  0
##   nc south   totcoll
## 1  0     0 0.0000000
## 2  1     0 7.0333333
## 3  0     0 0.0000000
## 4  0     0 0.2666667
## 5  0     1 0.0000000
## 6  0     1 0.0000000
```

```
#summary(data)
round(stat.desc(data, desc = TRUE, basic = TRUE), 2)
```

```
##                 female   phsrank      BA      AA   black hispanic
## nbr.val      6763.00   6763.00 6763.00 6763.00 6763.00  6763.00
## nbr.null     3249.00     12.00 4690.00 6465.00 6120.00  6446.00
## nbr.na          0.00      0.00    0.00    0.00    0.00     0.00
## min             0.00      0.00    0.00    0.00    0.00     0.00
## max             1.00     99.00    1.00    1.00    1.00     1.00
## range           1.00     99.00    1.00    1.00    1.00     1.00
## sum          3514.00 379790.00 2073.00  298.00  643.00   317.00
## median          1.00     50.00    0.00    0.00    0.00     0.00
## mean            0.52     56.16    0.31    0.04    0.10     0.05
## SE.mean         0.01      0.30    0.01    0.00    0.00     0.00
## CI.mean.0.95    0.01      0.58    0.01    0.00    0.01     0.01
## var             0.25    589.18    0.21    0.04    0.09     0.04
## std.dev         0.50     24.27    0.46    0.21    0.29     0.21
## coef.var        0.96      0.43    1.50    4.66    3.09     4.51
##                         id     exper      jc     univ    lwage  stotal
## nbr.val           6763.00   6763.00 6763.00  6763.00  6763.00 6763.00
## nbr.null             0.00      0.00 5110.00  3307.00     0.00 1528.00
## nbr.na               0.00      0.00    0.00     0.00     0.00    0.00
## min                 19.00      3.00    0.00     0.00     0.56   -3.32
## max              89958.00    166.00    3.83     7.50     3.91    2.24
## range            89939.00    163.00    3.83     7.50     3.36    5.56
## sum          274684136.00 827667.00 2291.94 13027.39 15203.87  321.13
## median           39301.00    129.00    0.00     0.20     2.28    0.00
## mean             40615.72    122.38    0.34     1.93     2.25    0.05
## SE.mean            303.76      0.41    0.01     0.03     0.01    0.01
## CI.mean.0.95       595.47      0.80    0.02     0.05     0.01    0.02
## var          624031994.37   1117.43    0.60     5.28     0.24    0.73
## std.dev          24980.63     33.43    0.77     2.30     0.49    0.85
## coef.var             0.62      0.27    2.28     1.19     0.22   17.98
##               smcity medcity  submed  lgcity   sublg vlgcity  subvlg
## nbr.val      6763.00 6763.00 6763.00 6763.00 6763.00 6763.00 6763.00
## nbr.null     4833.00 5969.00 6299.00 6124.00 6174.00 6367.00 6333.00
## nbr.na          0.00    0.00    0.00    0.00    0.00    0.00    0.00
## min             0.00    0.00    0.00    0.00    0.00    0.00    0.00
## max             1.00    1.00    1.00    1.00    1.00    1.00    1.00
## range           1.00    1.00    1.00    1.00    1.00    1.00    1.00
## sum          1930.00  794.00  464.00  639.00  589.00  396.00  430.00
## median          0.00    0.00    0.00    0.00    0.00    0.00    0.00
## mean            0.29    0.12    0.07    0.09    0.09    0.06    0.06
## SE.mean         0.01    0.00    0.00    0.00    0.00    0.00    0.00
## CI.mean.0.95    0.01    0.01    0.01    0.01    0.01    0.01    0.01
## var             0.20    0.10    0.06    0.09    0.08    0.06    0.06
## std.dev         0.45    0.32    0.25    0.29    0.28    0.23    0.24
## coef.var        1.58    2.74    3.68    3.10    3.24    4.01    3.84
##                   ne      nc   south  totcoll
## nbr.val      6763.00 6763.00 6763.00  6763.00
## nbr.null     5338.00 4742.00 4551.00  2483.00
## nbr.na          0.00    0.00    0.00     0.00
## min             0.00    0.00    0.00     0.00
## max             1.00    1.00    1.00    10.07
## range           1.00    1.00    1.00    10.07
## sum          1425.00 2021.00 2212.00 15319.34
## median          0.00    0.00    0.00     1.51
## mean            0.21    0.30    0.33     2.27
## SE.mean         0.00    0.01    0.01     0.03
## CI.mean.0.95    0.01    0.01    0.01     0.06
## var             0.17    0.21    0.22     5.43
## std.dev         0.41    0.46    0.47     2.33
## coef.var        1.94    1.53    1.43     1.03
```

There are 6763 observations of 23 variables. There are 0 `NA`s in the whole dataset.

One of the variables, `id`, is an ID number, so it should be unrelated with any other and hence of no interest. But it helps us to determine if the **random sampling** assumption (MRL.2) is met… which may not be the case: there are no observations for IDs between 65,500 and 70,000, and fewer members of the sample have an ID higher than 70,000, compared to lower values (see the missing ranges between 65,500 and 70,000, as well as the histogram, in the next page).

```
# Assign each ID to a 500-range
id_range = cut(data$id, breaks = seq(1, (ceiling(max(data$id)/500) + 1)*500,
                                      by = 500))
# Check unassigned ranges / levels
setdiff(levels(id_range), droplevels(id_range))
```

```
## [1] "(6.55e+04,6.6e+04]" "(6.6e+04,6.65e+04]" "(6.65e+04,6.7e+04]"
## [4] "(6.7e+04,6.75e+04]" "(6.75e+04,6.8e+04]" "(6.8e+04,6.85e+04]"
## [7] "(6.85e+04,6.9e+04]" "(6.9e+04,6.95e+04]" "(6.95e+04,7e+04]"
```

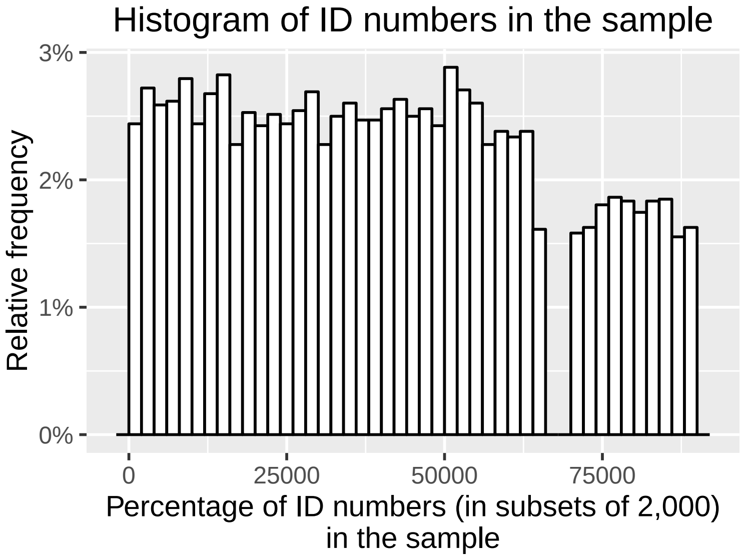

Figure 1: Histogram of ID numbers (in subsets of 2,000) in the sample

Without information about how the IDs were assigned, we will have to assume that for some reason those IDs between 65,500 and 70,000 did not even exist in the population, and that IDs higher than 70,000 have been assigned randomly—not subsequentially—and recently, and hence it is normal than fewer people in the sample have such higher IDs. I.e., we will assume that the sampling distribution resembles the distribution of the population and the dataset is a random sample of the population.

## 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:**

$$\begin{gathered}\log(\mathbf{wage}) = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{jc} + \beta_\mathbf{2}\mathbf{univ} + \beta_\mathbf{3}\mathbf{exper} + \beta_\mathbf{4}\mathbf{black} + \beta_\mathbf{5}\mathbf{hispanic} \\ + \beta_\mathbf{6}\mathbf{AA} + \beta_\mathbf{7}\mathbf{BA} + \beta_\mathbf{8}\mathbf{exper}\cdot\mathbf{black} + \mathbf{e}\end{gathered}$$

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

```r
# Set of independent variables
params <- c('jc', 'univ', 'exper', 'black', 'hispanic', 'AA', 'BA')
# Include interaction terms
params2 <- c(params, 'exper*black')
# Include dependent variable
var_of_interest <- c('lwage', params)
# (Reminder of) Meaning of each variable
subset(desc, variable %in% var_of_interest)
```

```
##    variable                          label
## 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
## 8     exper total (actual) work experience
## 9        jc           total 2-year credits
## 10     univ           total 4-year credits
## 11    lwage                log hourly wage
```

```r
model1 <- lm(as.formula(paste(var_of_interest[!var_of_interest %in% params],
                              paste(params2, sep = "", collapse = " + "),
                              sep = " ~ ")), data = data)
```

Table 1: Regression summary

| | *Dependent variable:* |
|---|---|
| | lwage |
| jc | 0.064*** |
| | (0.008) |
| univ | 0.073*** |
| | (0.003) |
| exper | 0.005*** |
| | (0.0002) |
| black | 0.033 |
| | (0.069) |
| hispanic | −0.019 |
| | (0.025) |
| AA | −0.008 |
| | (0.027) |
| BA | 0.018 |
| | (0.017) |
| exper:black | −0.001* |
| | (0.001) |
| Constant | 1.477*** |
| | (0.023) |
| F Statistic | 248.019*** |
| df | 8; 6754 |
| Observations | 6,763 |
| $R^2$ | 0.228 |
| Adjusted $R^2$ | 0.227 |
| Residual Std. Error | 0.429 |

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

## Question 3

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

## 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 university education is equal to the return to 1 year of working experience.**

## Question 6

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

## 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?**

## 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?**

Table 2: Table caption

| | uno | dos |
|---|---|---|
| x | **2.039**$^{***}$ | |
| | (0.028) | |
| z | | **0.556**$^{**}$ |
| | | (0.196) |
| (Intercept) | 0.014 | $-0.223$ |
| | (0.024) | (0.195) |
| $R^2$ | 0.985 | 0.075 |
| $R^2_{adj}$ | 0.985 | 0.066 |
| F | 5179.804 | 8.081 |
| N | 100 | 100 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 3: Table

| | Model 1 | Model 2 |
|---|---|---|
| (Intercept) | 0.014 | $-0.223$ |
| | (0.024) | (0.193) |
| x | **2.039**$^{***}$ | |
| | (0.025) | |
| z | | **0.556**$^{**}$ |
| | | (0.197) |
| $R^2$ | 0.985 | 0.075 |
| Adj. $R^2$ | 0.985 | 0.066 |
| Num. obs. | 100 | 100 |
| RMSE | 0.244 | 1.927 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

...

...

Table 4: test stargzer

| | YY | |
|---|---|---|
| | (1) | (2) |
| XX | 2.039*** | |
| | (0.028) | |
| ZZ | | 0.556** |
| | | (0.196) |
| (Intercept) | 0.014 | −0.223 |
| | (0.024) | (0.195) |
| F Statistic | 5,179.804*** | 8.081** |
| df | 1; 98 | 1; 98 |
| Observations | 100 | 100 |
| $R^2$ | 0.985 | 0.075 |
| Adjusted $R^2$ | 0.985 | 0.066 |
| Residual Std. Error | 0.244 | 1.927 |

·p<0.1; *p<0.05; **p<0.01; ***p<0.001