W271-2 – Spring 2016 – **HW 3**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

February 17, 2016

Contents

Exercises

Question 1

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

```
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
```

```
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
```

```r
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
```

```r
#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).

```r
# 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]"
```

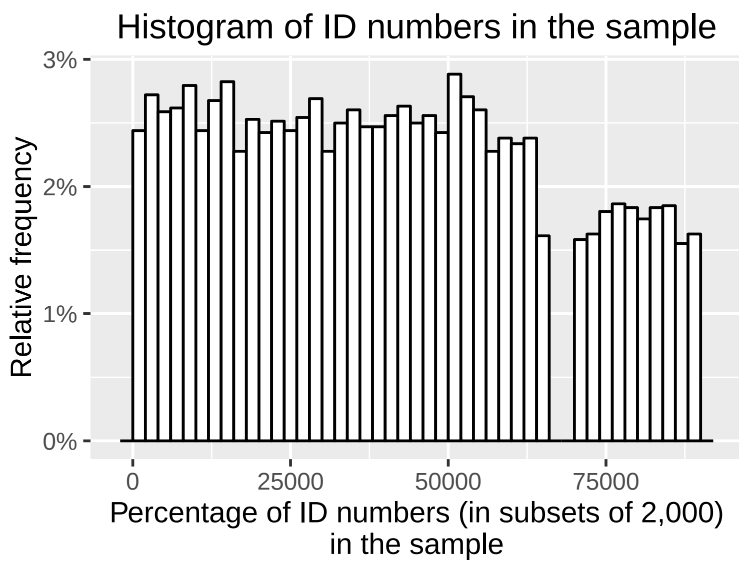

Figure 1: Histogram of ID numbers (in subsets of 2,000) in the sample

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:

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

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

Before estimating the regression, let's remember the meaning and summary statistics of the variables of interest:

```r
# Set of independent variables
params <- c('jc', 'univ', 'exper', 'black', 'hispanic', 'AA', 'BA')
# Include interaction terms
params_plus_interaction <- c(params, 'exper*black')
# Include dependent variable
vars_of_interest <- c('lwage', params)
# (Reminder of) Meaning of each variable
subset(desc, variable %in% vars_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
# Summary of variables of interest only
summary(data[, vars_of_interest])
```

```
##      lwage              jc              univ           exper      
##  Min.   :0.5555   Min.   :0.0000   Min.   :0.000   Min.   :  3.0  
##  1st Qu.:1.9253   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:104.0  
##  Median :2.2763   Median :0.0000   Median :0.200   Median :129.0  
##  Mean   :2.2481   Mean   :0.3389   Mean   :1.926   Mean   :122.4  
##  3rd Qu.:2.5969   3rd Qu.:0.0000   3rd Qu.:4.200   3rd Qu.:149.0  
##  Max.   :3.9120   Max.   :3.8333   Max.   :7.500   Max.   :166.0  
##      black            hispanic             AA                BA        
##  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   :0.09508   Mean   :0.04687   Mean   :0.04406   Mean   :0.3065  
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000  
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000  
```

```r
# OLS model
model1 <- lm(as.formula(paste(vars_of_interest[!vars_of_interest %in% params],
                              paste(params_plus_interaction, sep = "",
                                    collapse = " + "), sep = " ~ ")),
             data = data)
```

Table 1: Regression summary

	Dependent variable:
	lwage
Junior college (2-yr credits)	0.0638***
	(0.0076)
University (4-yr credits)	0.0733***
	(0.0034)
Work experience (months)	0.0050***
	(0.0002)
Black	0.0332
	(0.0687)
Experience * Black	−0.0013*
	(0.0005)
Hispanic	−0.0194
	(0.0250)
Associate's degree	−0.0078
	(0.0275)
Bachelor's degree	0.0177
	(0.0166)
Intercept (Constant)	1.4773***
	(0.0229)
F Statistic	248.019***
df	8; 6754
Observations	6,763
R^2	0.2282
Adjusted R^2	0.2272
Residual Std. Error	0.4287

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

All the p and F values shown in this document have been estimated using heteroskedasticity-robust standard errors.

As shown in Table 1 above, the coefficients of interest are $\hat{\beta}_4 = 0.0332$ (0.0687) and $\hat{\beta}_8 = -0.0013$ (0.0005), respectively. They mean that, all other factors held constant *at their baseline values* (i.e., 0 total 2- and 4-year credits, not being Hispanic, having no experience and neither a Bachelor's nor an Associate's degree):

- on average, being African-American increases the log of hourly wage by 0.0332. I.e., on average an African-American earns about 3.3% more than a person that is not. The estimate of this coefficient is not statistically significant ($p = 0.629$).
- on average, each additional month of experience (the maximum value of `exper` is 166 so we assume that its units of measurement are months, not years) decreases the log of wage by -0.0013; i.e., roughly a 0.1% decrease. The estimate of this coefficient is statistically significant at the 0.05 level.

The diagnostic plots (below) show that all the CLM assumptions are met to some extent (the residuals are not totally homoskedastic—we will analyze this point in Question 8):

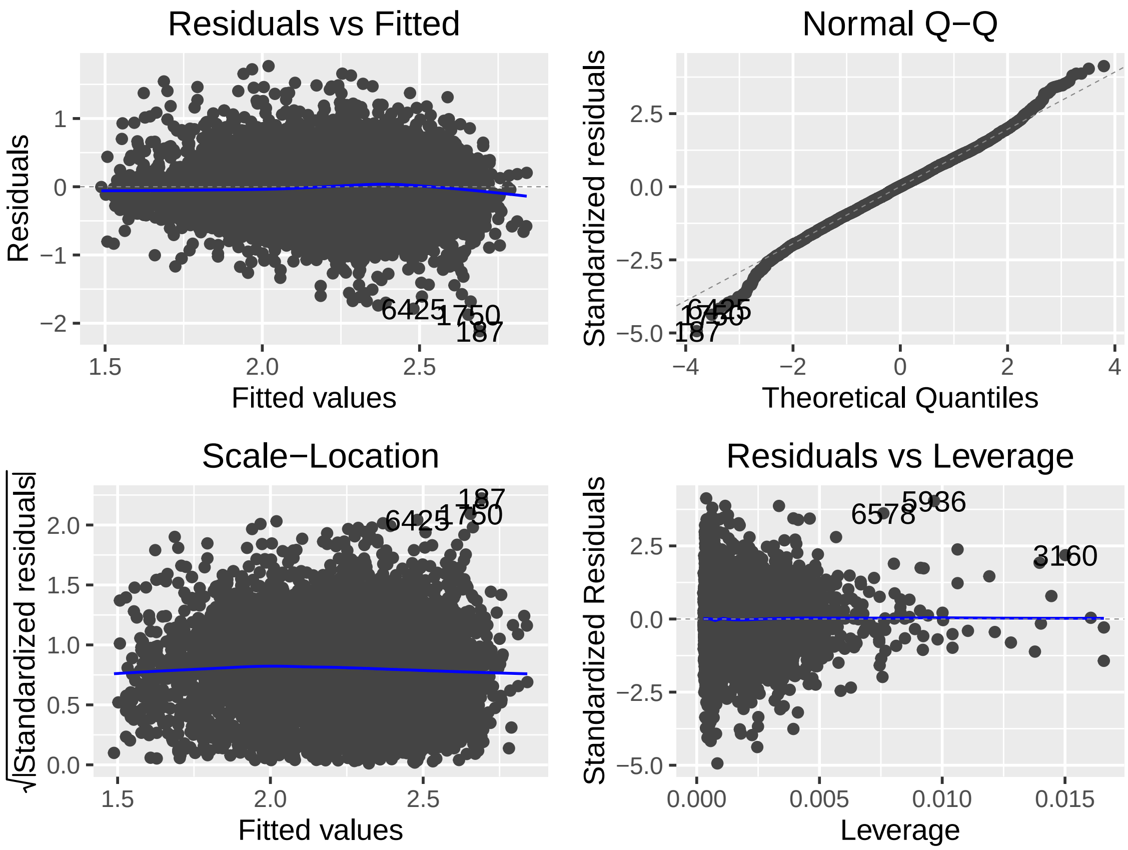

Figure 2: Diagnotic plots of the regression model

Anyway, using `exper` = 0 as the baseline does not make too much sense (the minimum value that `exper` takes is 3). We can build a new model where the mean value of that variable (122.4, presumably about 10.2 years if the former value is in months) is the baseline by replacing `exper` by its value centered around its mean , `exper_mean = exper - mean(exper)`.

```
# Center exper around its mean
data2 <- data[, vars_of_interest]
data2$exper_mean <- data2$exper - mean(data2$exper)
params2 <- gsub('exper', 'exper_mean', params)
params_plus_interaction2 <- gsub('exper', 'exper_mean',
                                 params_plus_interaction)
vars_of_interest2 <- c('lwage', params2)
model2 <- lm(as.formula(paste(vars_of_interest2[!vars_of_interest2 %in%
                                                  params2],
                              paste(params_plus_interaction2, sep = "",
                                    collapse = " + "), sep = " ~ ")),
             data = data2)
```

As shown in Table 2 in the following page, all the coefficients except the intercept and the one corresponding to `black` remain the same. Now $\hat{\beta}_4$ becomes -0.122 (0.018), and it's highly statistically significant ($p = 7.4e-12$). I.e., all other factors held constant *at their (new) baseline values* (i.e., 0 total 2- and 4-year credits, having neither a Bachelor's nor an Associate's degree, and 10.2 years of experience), being African-American decreases the log of hourly wage by 0.122 on average (about a 12.2% decrease). An additional month of experience still reduces the hourly wage by roughly another 0.1 percentage points.

Table 2: Regression summary using 0 and its mean (122.4) as the baselines values of exper

	Dependent variable:	
	lwage	lwage
	(1)	(2)
Junior college (2-yr credits)	0.0638***	0.0638***
	(0.0076)	(0.0076)
University (4-yr credits)	0.0733***	0.0733***
	(0.0034)	(0.0034)
Work experience (months)	0.0050***	
	(0.0002)	
Work experience (months) with respect to mean (122.4)		0.0050***
		(0.0002)
Black	0.0332	−0.1220***
	(0.0687)	(0.0178)
Experience * Black	−0.0013*	
	(0.0005)	
Experience with respect to mean * Black		−0.0013*
		(0.0005)
Hispanic	−0.0194	−0.0194
	(0.0250)	(0.0250)
Associate's degree	−0.0078	−0.0078
	(0.0275)	(0.0275)
Bachelor's degree	0.0177	0.0177
	(0.0166)	(0.0166)
Intercept (Constant)	1.4773***	2.0921***
	(0.0229)	(0.0079)
F Statistic	248.019***	248.019***
df	8; 6754	8; 6754
Observations	6,763	6,763
R^2	0.2282	0.2282
Adjusted R^2	0.2272	0.2272
Residual Std. Error	0.4287	0.4287

·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%.

Since we are using the log of the dependent variable, a coefficient of 7% can be approximated by 0.07. But, strictly speaking, the coefficient $\hat{\beta}_2$ should be $\log(0.07 + 1) = 0.0677$ to have a return to university education of exactly 7% (because $e^{\log(0.07+1)} - 1 = 0.07$). We will test both hypotheses and see that we fail to reject the former (the approximation) but the latter (the exact hypothesis) is rejected at the 0.1 level (the size of our sample has a large influence, so overall we could say we should not reject the hypothesis, either approximate or exact).

Because we want to test $H_0 : \beta_2 = \mu$ (where $\mu = 0.07$ or 0.0677), the t statistic we have to use is:

$$t = \frac{\hat{\beta}_2 - \mu}{se(\hat{\beta}_2)}$$

We can use the values for $\hat{\beta}_2$ and $se(\hat{\beta}_2)$ that we obtained before (shown in Table 1) to estimate the p value:

```
# New t statististic
(t <- (coeftest(model1, vcovHC(model1))[2+1, 1] - .07) /
   coeftest(model1, vcovHC(model1))[2+1, 2])
```

```
## [1] 0.9746433
```

```
(t_exact <- (coeftest(model1, vcovHC(model1))[2+1, 1] - log(0.07 + 1)) /
  coeftest(model1, vcovHC(model1))[2+1, 2])
```

```
## [1] 1.670235
```

```
# New p value
2*pt(t, dim(data)[1] - 1, lower.tail = FALSE)
```

```
## [1] 0.3297721
```

```
2*pt(t_exact, dim(data)[1] - 1, lower.tail = FALSE)
```

```
## [1] 0.09491913
```

The t statistic for $\beta_2 = 0.07$ is quite low (about half the typical critical value of 1.96 for a 0.05 significance level), so the probability of obtaining such a low value is quite high (about 33%) and hence we fail to reject the null hypothesis: we do not have enough evidence to believe that the return to university education is different than 7%.

But the t statistic for $\beta_2 = \log(0.07 + 1) = 0.0677$ is below 1.64, the critical value for a 0.1 significance level (when the number of observations is huge and we can approximate a t distribution by the standard normal distribution), so the probability of obtaining such a low value is below 0.1 so we reject the hypothesis at that significance level.

> Another way to test the hypothesis would be the following: suppose that our (simplified) model is $y = \beta_0 + \beta_2 x$; testing $\beta_2 = \mu$ is equivalent to test $\beta_2' = 0$, where $\beta_2' = \beta_2 - \mu$. Replacing β_2 by $\beta_2' + \mu$ we can rewrite the model as $y' = y - \mu x = \beta_0 + \beta_2' x$. If we do so, we should get an estimate $\hat{\beta}_2'$ close to zero, with the same t statistic and p value that we obtained above:

```r
data3 <- data[, vars_of_interest]
data3$lwage_minus7univ <- data3$lwage - .07 * data3$univ
vars_of_interest3 <- c('lwage_minus7univ', params)
model3 <- lm(as.formula(paste(vars_of_interest3[!vars_of_interest3 %in%
                                                   params],
                              paste(params_plus_interaction, sep = "",
                                    collapse = " + "), sep = " ~ ")),
             data = data3)
coeftest(model3, vcovHC(model3))
```

```
##
## 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         0.00328063  0.00336598  0.9746  0.32977
## 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
## exper:black -0.00126790  0.00053779 -2.3576  0.01842 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
data4 <- data[, vars_of_interest]
data4$lwage_minus7univ <- data4$lwage - log(0.07 + 1) * data4$univ
vars_of_interest4 <- c('lwage_minus7univ', params)
model4 <- lm(as.formula(paste(vars_of_interest3[!vars_of_interest4 %in%
                                                   params],
                              paste(params_plus_interaction, sep = "",
                                    collapse = " + "), sep = " ~ ")),
             data = data4)
coeftest(model4, vcovHC(model4))[3, ]
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## 0.005621986 0.003365985 1.670235126 0.094919189
```

And yet another way would be:

```r
linearHypothesis(model1, c("univ = 0.07"), vcov = vcovHC(model1))$'Pr(>F)'[2]
```

```
## [1] 0.3297721
```

```r
linearHypothesis(model1, c(paste("univ = ", log(0.07 + 1))),
                 vcov = vcovHC(model1))$'Pr(>F)'[2]
```

```
## [1] 0.09491919
```

Question 4

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

Without including the corresponding interaction term (`jc * black`), the most we can test is whether the partial effect of junior college education has the same intercept for black and non-black (after controlling for all other factors), but no test about different slopes can be carried.

```
linearHypothesis(model1, c("black = 0"), vcov = vcovHC(model1))
```

```
## 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
```

This would be equivalent to test the effect of dropping `black` from the regression model, so the p value is equal to the one of the t statistic of 'black:

```
coeftest(model1, vcov = vcovHC(model1))[4+1, 4]
```

```
## [1] 0.6293633
```

The p value is quite high, so we fail to reject the hypothesis.

A visual inspection of `jc` against `lwage` for both groups seems to suggest that the intercepts are different (and the slopes are almost the same)… though the test is not significant.

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  1.47674248  0.02293351 64.3924  < 2e-16 ***
## jc           0.06590812  0.00769696  8.5629  < 2e-16 ***
## univ         0.07334067  0.00336476 21.7967  < 2e-16 ***
## exper        0.00502224  0.00016841 29.8222  < 2e-16 ***
## black        0.04287092  0.06983552  0.6139  0.53931
## hispanic    -0.01945976  0.02499028 -0.7787  0.43619
## AA          -0.00876144  0.02746021 -0.3191  0.74969
## BA           0.01742583  0.01656112  1.0522  0.29274
## exper:black -0.00128651  0.00054186 -2.3743  0.01761 *
## jc:black    -0.03373825  0.03322134 -1.0156  0.30988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

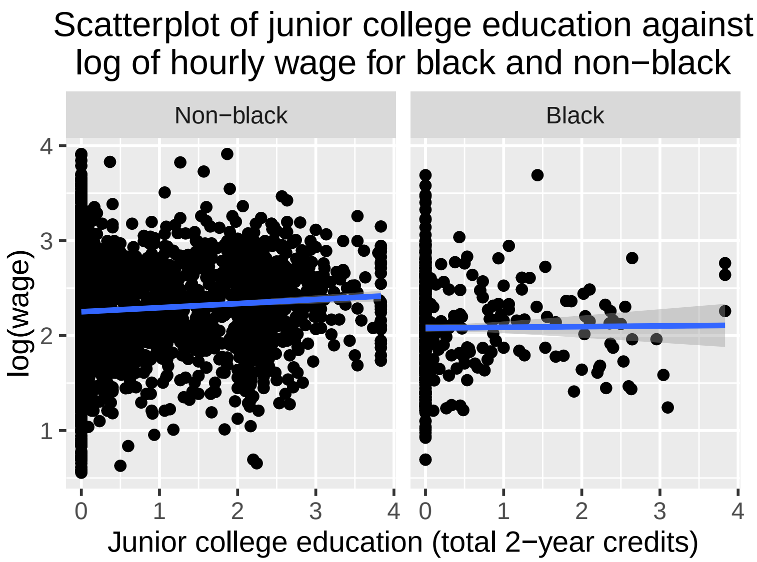

Figure 3: Scatterplot of junior college education against log of hourly wage for black and non-black

```
linearHypothesis(model5, c("black = 0", "jc:black = 0", "exper:black = 0"),
                 vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## black = 0
## jc:black = 0
## exper:black = 0
##
## Model 1: restricted model
## Model 2: lwage ~ jc + univ + exper + black + hispanic + AA + BA + exper *
##     black + jc * black
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   6756
## 2   6753  3 18.135 1.022e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# m <- lm(lwage ~ jc + univ + exper + hispanic + AA + BA, data)
# m1 <- lm(lwage ~ jc + univ + exper + hispanic + AA + BA,
#          data[data$black == 0, ])
# m2 <- lm(lwage ~ jc + univ + exper + hispanic + AA + BA,
#          data[data$black == 1, ])
# f <- (sum(m$residuals^2) - sum(m1$residuals^2) - sum(m2$residuals^2)) /
#   (sum(m1$residuals^2) + sum(m2$residuals^2)) * (dim(data)[1] - 2*3) / (2 + 1)
# pf(f, 3, dim(data)[1] - 4, lower.tail = FALSE)
```

Question 5

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

We want to test the hypothesis $H_0 : \hat{\beta}_{univ} = 12 \cdot \hat{\beta}_{exper}$:

```
linearHypothesis(model1, c("univ = 12*exper"), vcov = vcovHC)
```

```
## 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
```

So we reject that hypothesis at the 0.001 level.

> Another way would be to rewrite the model $log(wage) = \beta_0 + \beta_2 univ + \beta_3 exper + \ldots$ as $log(wage) = \beta_0 + \beta_3(12 \cdot univ + expr) + \beta_3' univ + \ldots$ (replacing β_2 by $12 \cdot \beta_3 + \beta_3'$) and check whether β_3' is statistically significantly different from zero:

```
model6 <- lm(lwage ~ jc + I(12*univ + exper) + univ + black + hispanic + AA +
               BA + exper * black, data)
coeftest(model6, vcovHC(model6))[1:4, ]
```

```
##                       Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)        1.477331546 0.0229351203 64.413508  0.000000e+00
## jc                 0.063792607 0.0076120765  8.380448  6.353465e-17
## I(12 * univ + exper) 0.005023413 0.0001684049 29.829371 8.860458e-184
## univ               0.012999678 0.0037576755  3.459500  5.445231e-04
```

Interestingly, we fail to reject the two null hypotheses that the return to junior college education is equal to the return to:

1. 1 year of working experience
2. university education

```r
linearHypothesis(model1, c("jc = 12*exper"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## jc - 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 0.1956 0.6583
```

```r
linearHypothesis(model1, c("univ = jc"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## - jc  + univ = 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 1.538  0.215
```

Question 6

Test the overall significance of this regression.

The value of the F statistic for the overall significance of the regression ($F = 248.019, p = 0$) was already shown in Table 1, Question 2: we reject the hypothesis that none of the explanatory variables included in this regression model help to explain `lwage` (i.e., at least one of the coefficients is different from zero).

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?

```
params_plus_interaction_square <- c(params_plus_interaction, 'I(exper^2)')
model7 <- lm(as.formula(paste(vars_of_interest[!vars_of_interest %in% params],
                              paste(params_plus_interaction_square, sep = "",
                                    collapse = " + "), sep = " ~ ")),
             data = data)
100 * (coeftest(model7, vcov = vcovHC(model7))[3+1, 1] +
         2*coeftest(model7, vcov = vcovHC(model7))[8+1, 1] *
         mean(data$exper)) * 12
```

```
## [1] 6.15343
```

```
100 * coeftest(model1, vcov = vcovHC(model1))[3+1, 1] * 12
```

```
## [1] 6.028096
```

As shown in Table 3 in the following page, $\hat{\beta}_{exper}$ has now decreased to 0.0043 (and it's still highly statistically significant) and '$\hat{\beta}_{exper^2} = 0.000003$ (not significant, \$p = 0.38). This means that, holding all other variables fixed, the estimated return to work experience can be approximated by:

$$\Delta\widehat{lwage} \simeq \left(\hat{\beta}_{exper} + 2\hat{\beta}_{exper^2} exper\right) \Delta exper$$

$$\%\Delta\widehat{wage} \simeq 100 \left(\hat{\beta}_{exper} + 2\hat{\beta}_{exper^2} exper\right) \Delta exper$$

I.e., $\%\boldsymbol{\Delta}\widehat{\mathbf{wage}} \simeq (\mathbf{0.430081} + \mathbf{0.00067579} \cdot \mathbf{exper})\boldsymbol{\Delta}\mathbf{exper}$. A 1-year increase in work experience for someone with an average experience will increase his or her hourly wage by 6.15% (compared to an average 6.03% increase if we don't include the square term of working experience in the model; but keep in mind the contribution of that effect is uncertain due to its lack of signficance).

Table 3: Regression summary with and without including the square term of working experience

	Dependent variable:	
	lwage	lwage
	(1)	(2)
Junior college (2-yr credits)	0.0642***	0.0638***
	(0.0076)	(0.0076)
University (4-yr credits)	0.0738***	0.0733***
	(0.0035)	(0.0034)
Work experience (months)	0.0043***	0.0050***
	(0.0008)	(0.0002)
Work experience^2	0.000003	
	(0.000004)	
Black	0.0299	0.0332
	(0.0684)	(0.0687)
Experience * Black	−0.0012*	−0.0013*
	(0.0005)	(0.0005)
Hispanic	−0.0193	−0.0194
	(0.0250)	(0.0250)
Associate's degree	−0.0075	−0.0078
	(0.0275)	(0.0275)
Bachelor's degree	0.0180	0.0177
	(0.0166)	(0.0166)
Intercept (Constant)	1.5101***	1.4773***
	(0.0436)	(0.0229)
F Statistic	220.039***	248.019***
df	9; 6753	8; 6754
Observations	6,763	6,763
R^2	0.2282	0.2282
Adjusted R^2	0.2272	0.2272
Residual Std. Error	0.4287	0.4287

·p<0.1; *p<0.05; **p<0.01; ***p<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?

We can start running run a Breusch-Pagan test to test for heteroskedasticity. The results of two implementations of such test in R are slightly different, but both are significant, which indicates heteroskedasticity . . . but our sample is so large (6763 observations) that this test is very likely to yield a significant result even in the absence of heteroskedasticity.

```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 39.425, df = 8, p-value = 4.099e-06
```

```
ncvTest(model1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 3.9443    Df = 1     p = 0.04703041
```

For such a huge dataset, it's better to use diagnostic plots. Let's plot again two of the graphs already shown in Figure 2, this time with more detail:

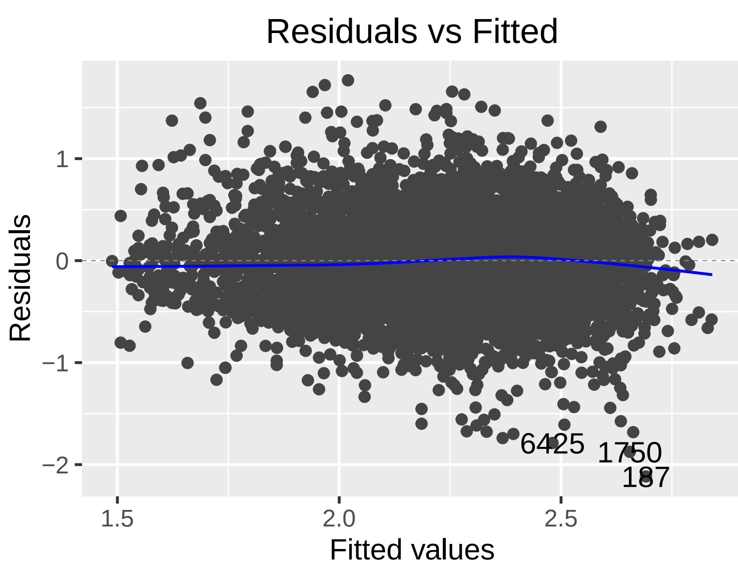

Figure 4: Residuals vs. Fitted values plot

The thickness of the band in the residuals vs. fitted plot is almost the same for all fitted values. There seems to be less variance on both extremes, but that might be due to a lack of observations for the most extreme values of `lwage`.

Since we are presented with a hypothesis that involves `univ`, let's also plot how the residuals vary with that specific variable. As shown in Figure 5, the variance is higher for extreme and middle values of `univ` (around the minimum, mean, and maximum) and lower (though not too much) for values inbetween.

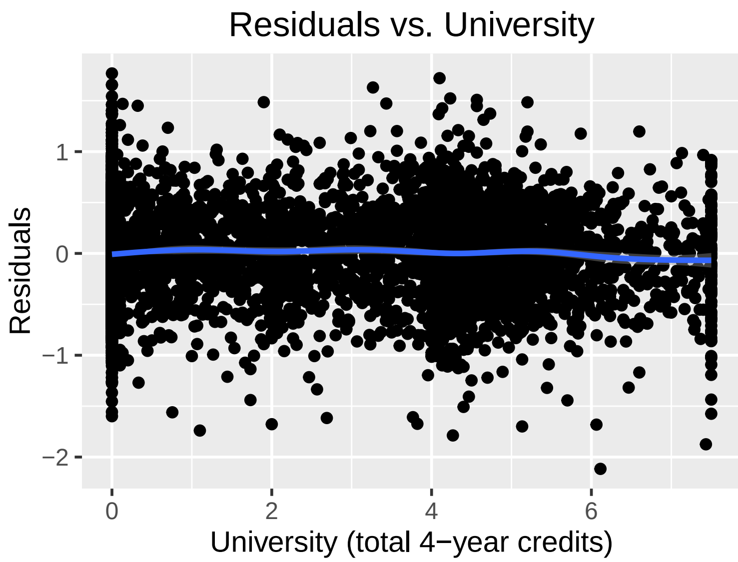

Figure 5: Residuals vs. University

Finally, the scale-location plot shows a almost horizontal band of points (which does not go upwards or downwards), which also suggests homoskedasticity.

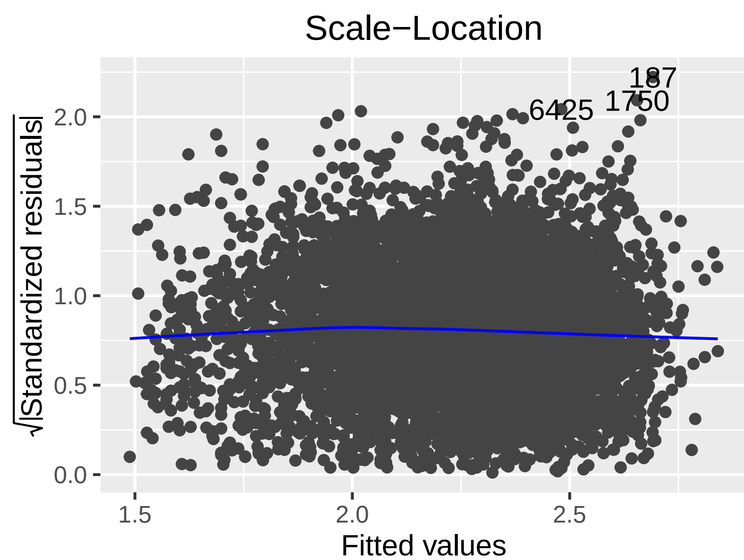

Figure 6: Scale-Location plot

The assumption of homoskedasticity (together with assumptions MLR.1 through MLR.4) is required for the OLS estimators to be BLUE (the best—the ones with less variance among—linear unbiased estimator). It is also required (together with the normality of errors) to ensure that the distribution of the t and F statistics follow t and F distributions, respectively. So if this assumption were broken, we would not be able to estimate the significance level of a hypothesis test like that university education has no effect on salary.

There are some signs that the assumption is broken, which we can overcome by using **heteroskedasticity-robust standard errors** . . . which we have done through this whole document. This way, we can be confident in the significance results of testing $H_0 : \beta_{univ} = 0$ (which we reject at the 1e-101 significance level!).