MIDS-W271-4-HW3

Homework 3

Question 1

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

---

```r
library(car)
library(lmtest)
library(sandwich)

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 deg
## - attr(*, "version")= int 10
```

```
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
```

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:

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + \beta_4 black + \beta_5 hispanic + \beta_6 AA + \beta_7 BA + \beta_8 exper * black + \epsilon$$

Interpret the coefficents $\hat{\beta}_4$ and $\hat{\beta}_8$

Constructing our model we have:

```r
model1 <- lm(lwage ~ jc+univ+exper+black+hispanic+AA+BA+exper:black, data=data)

summary(model1)
```

```
##
## 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
```

$\hat{\beta}_4$ is an indicator variable to signify a member of a group. Since the only two race indicators are black and hispanic one could reason that the base group consists of all other races.

$\hat{\beta}_8$ is an interaction term used to explore the partial wage difference of job-experienced blacks compared to other races, controlling for the other terms in the regression model.

Question 3

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

$$H_0 : \beta_2 = 0.07$$
$$H_1 : \beta_2 \neq 0.07$$

$$t = (\hat{\beta}_2 - 0.07)/se(\hat{\beta}_2)$$
$$t = (0.0732806 - 0.07)/0.0031486 = 1.041923$$

We can calculate the p-value from the t-statistic as:

```r
pt((0.0732806 - 0.07)/0.0031486, 6793-9)
```

```
## [1] 0.8512578
```

This clearly indicates we can not reject the null hypothesis that $\beta_2 = 0.07$

Using the linearHypothesis function in R with robust heteroskedasticity standard errors confirms the result.

```r
linearHypothesis(model1, "univ=0.07", vcov=vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## univ = 0.07
##
## 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 0.9499 0.3298
```

Question 4

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

We can add an interaction term between junior college and black to the model:

```
model2 <- lm(lwage ~ jc+univ+exper+black+hispanic+AA+BA+black:exper+black:jc, data=data)

summary(model2)
```

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper + black + hispanic + AA +
##     BA + black:exper + black:jc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11547 -0.27839  0.00394  0.28669  1.76883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.4767425  0.0223831  65.976  < 2e-16 ***
## jc           0.0659081  0.0081083   8.128 5.13e-16 ***
## univ         0.0733407  0.0031490  23.290  < 2e-16 ***
## exper        0.0050222  0.0001667  30.134  < 2e-16 ***
## black        0.0428709  0.0619565   0.692    0.489
## hispanic    -0.0194598  0.0248909  -0.782    0.434
## AA          -0.0087614  0.0295610  -0.296    0.767
## BA           0.0174258  0.0156563   1.113    0.266
## exper:black -0.0012865  0.0004993  -2.577    0.010 **
## jc:black    -0.0337383  0.0289025  -1.167    0.243
## ---
## 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:  0.2283, Adjusted R-squared:  0.2273
## F-statistic:   222 on 9 and 6753 DF,  p-value: < 2.2e-16
```

The return to junior college is 0.0659 for non-blacks; for blacks the return to junior college is

```
0.06591 - 0.03374
```

```
## [1] 0.03217
```

A 3% differential, which is not economically large and it is not statistically significant: p = .243.

Question 5

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

$$H_0 : \beta_2 = \beta_3 + 1$$
$$t = \frac{\hat{\beta}_2 - (\hat{\beta}_3 + 1)}{se(\hat{\beta}_2 - \hat{\beta}_3)}$$

From the original regression equation: $log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + \beta_4 black + \beta_5 hispanic + \beta_6 AA + \beta_7 BA + \beta_8 exper * black + \epsilon$

We define a new variable as: $\theta_1 = \beta_2 - (\beta_3 + 1) = \beta_2 - \beta_3 - 1$

Rearranging we have: $\beta_2 = \theta_1 + beta_3 - 1$

Substituting for β_2 we can write:

$log(wage) = \beta_0 + \beta_1 jc + (\theta_1 + beta_3 - 1) univ + \beta_3 exper + \beta_4 black + \beta_5 hispanic + \beta_6 AA + \beta_7 BA + \beta_8 exper * black + \epsilon$

Multiplying and collecting terms yields:

$log(wage) = \beta_0 + \beta_1 jc + \theta_1 univ + beta_3 (univ + exper) + univ + \beta_3 exper + \beta_4 black + \beta_5 hispanic + \beta_6 AA + \beta_7 BA + \beta_8 exper * black + \epsilon$

One thing that we notice is that the intercept increases for every year of university. We also have a new coefficient of univ+exper. However, these numbers have different scales, so we should create normalized versions of them to add together.

```r
data$unexp <- scale(data$univ) + scale(data$exper)
model4 <- lm(lwage ~ jc+unexp+exper+black+hispanic+AA+BA+black:exper+black:jc, data=data)

summary(model4)
```

```
##
## Call:
## lm(formula = lwage ~ jc + unexp + exper + black + hispanic +
##     AA + BA + black:exper + black:jc, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11547 -0.27839  0.00394  0.28669  1.76883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.235e+00  3.488e-02  64.076  < 2e-16 ***
## jc           6.591e-02  8.108e-03   8.128 5.13e-16 ***
## unexp        1.685e-01  7.233e-03  23.290  < 2e-16 ***
## exper       -1.736e-05  2.608e-04  -0.067    0.947
## black        4.287e-02  6.196e-02   0.692    0.489
## hispanic    -1.946e-02  2.489e-02  -0.782    0.434
## AA          -8.761e-03  2.956e-02  -0.296    0.767
## BA           1.743e-02  1.566e-02   1.113    0.266
## exper:black -1.287e-03  4.993e-04  -2.577    0.010 **
## jc:black    -3.374e-02  2.890e-02  -1.167    0.243
## ---
## 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:  0.2283, Adjusted R-squared:  0.2273
## F-statistic:   222 on 9 and 6753 DF,  p-value: < 2.2e-16
```

Using the coefficient on exper of -1.736e-05 and se = 2.608e-04 we arrive at a t statistic of

```r
-1.736e-05/2.608e-04
```

```
## [1] -0.06656442
```

which is very small and therefore we do not reject the null that 1 year of experience is the same as the return on university.

Question 6

Test the overall significance of this regression.

The F-statistic shows highly significant, but we run a Wald test to accomodate for robust errors:

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

```
## Wald test
##
## Model 1: lwage ~ jc + univ + exper + black + hispanic + AA + BA + exper:black
## Model 2: lwage ~ 1
##   Res.Df Df      F    Pr(>F)
## 1   6754
## 2   6762 -8 248.02 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-test again shows as highly 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?

```r
data$exper2 <- data$exper^2
model3 <- lm(lwage ~ jc+univ+exper+exper2+black+hispanic+AA+BA+black:exper, data=data)

summary(model3)
```

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper + exper2 + black + hispanic +
##     AA + BA + black:exper, 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
```

The quadratic term for experience is slightly positive but not statistically significant. It implies a slightly increasing return to experience. The inflection point of the curve is given by

```r
4.3013e-03/(2*3.379e-06)
```

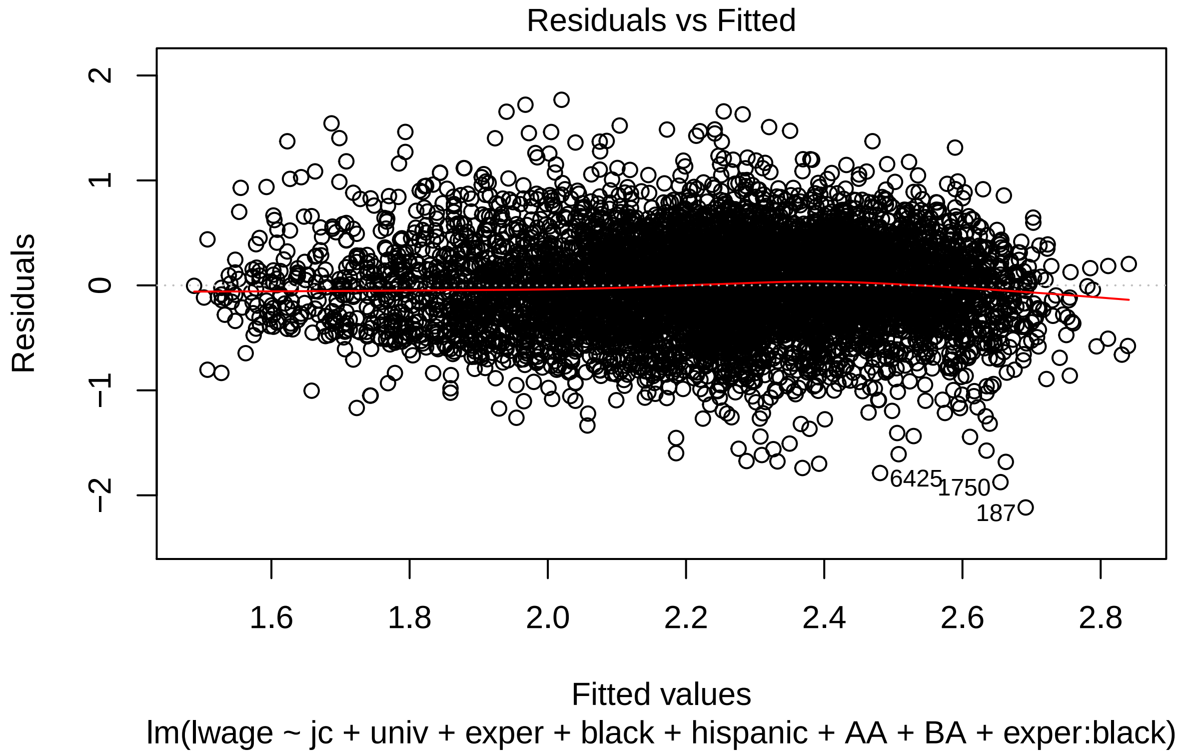
```
## [1] 636.4753
```

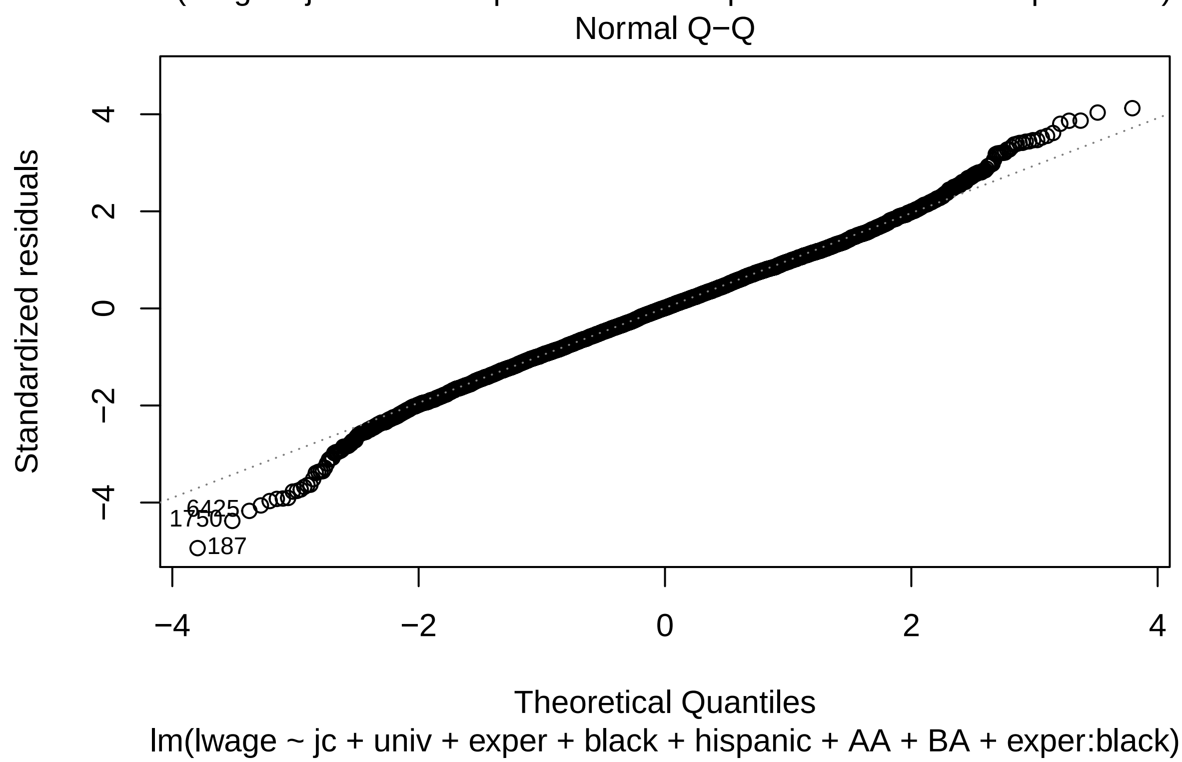
Which is very large at 636.5 and seems to indicate a very slightly increasing "lift" from experience as experience accumulates.

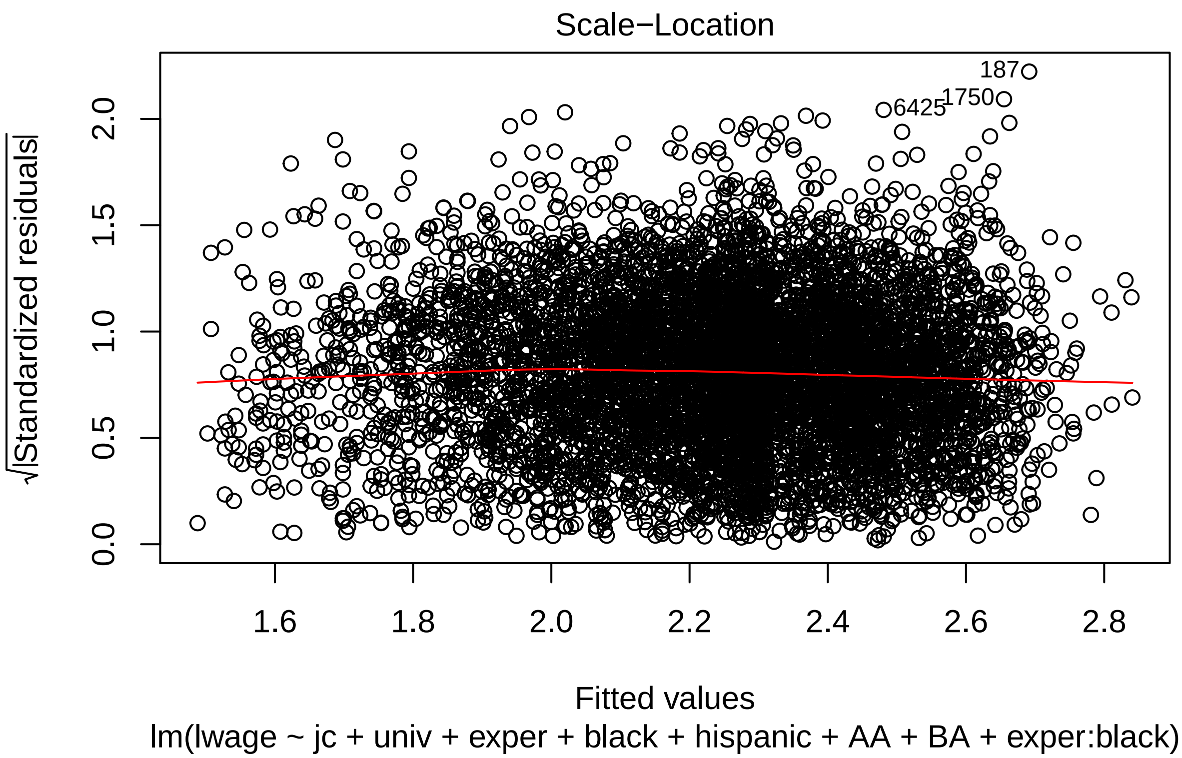
Question 8

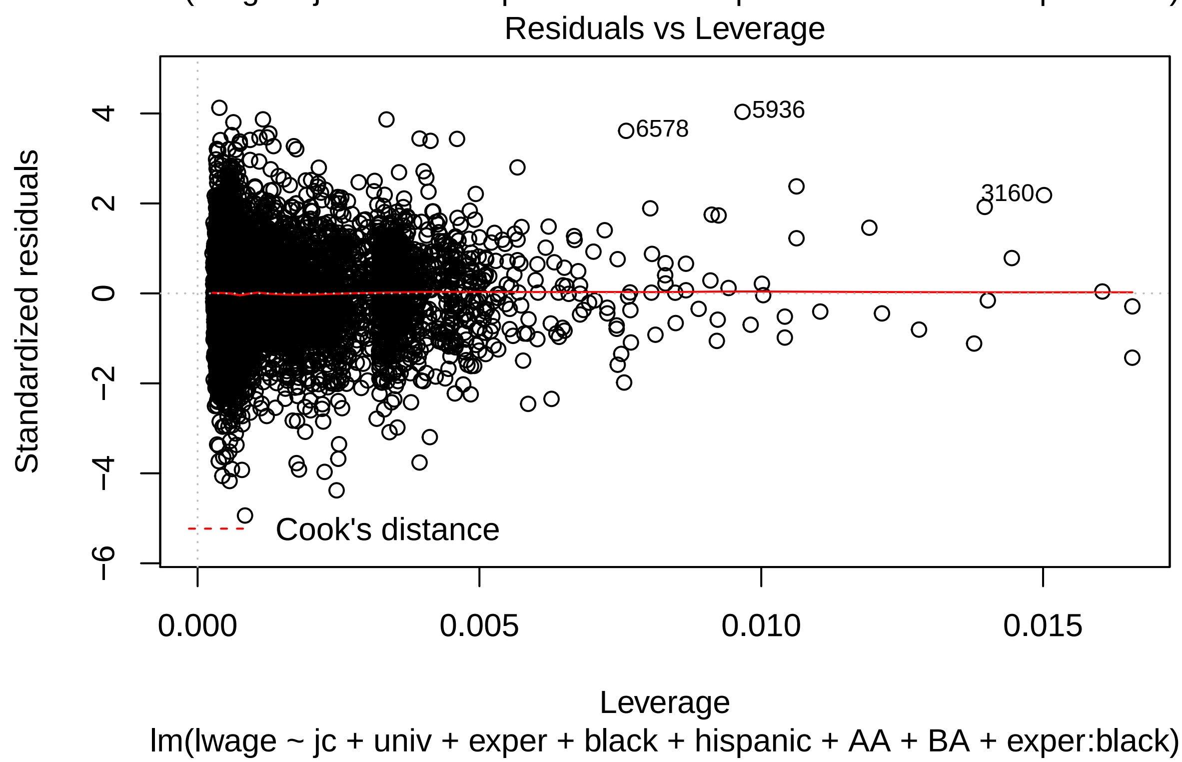
Provide the diagnosis of the homoskedasticity assumption. Does this assumption hold? If so, how does it affect the testing of no effect of university education on salary change? If not, what potential remedies are available?

Checking the diagnosic plots of our model we have the following plots. The residuals plot shows that the error is not uniform from left to right. The profile resembles the cross section of an airplane wing. However, the spline curve is mostly flat. We don't have zero conditional mean in this case. However, we do have over 6700 samples for which we can assume things to be asymptotically normal.


Residuals vs Fitted
Residuals
2
1
0
−1
−2
6425
1750
187
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 + exper:black)


Normal Q−Q
Standardized residuals
4
2
0
−2
−4
6425
1750
187
−4
−2
0
2
4
Theoretical Quantiles
lm(lwage ~ jc + univ + exper + black + hispanic + AA + BA + exper:black)


Scale−Location
√|Standardized residuals|
187
6425
1750
Fitted values
lm(lwage ~ jc + univ + exper + black + hispanic + AA + BA + exper:black)


Residuals vs Leverage
Standardized residuals
6578
5936
3160
Cook's distance
Leverage
lm(lwage ~ jc + univ + exper + black + hispanic + AA + BA + exper:black)