Problem Set 1 - Week 38, 2013

Problem 1

Consider the linear regression model

$$y_i = \beta_1 + f_i\beta_2 + \epsilon_i,$$

where y_i is the height of individual i, and f_i is a gender dummy, which takes values $f_i = 1$ for females and $f_i = 0$ for males. We observe n_f females and n_m males. The total sample size is $n = n_f + n_m$. Let $\bar{y}_f$ be the average of y_i in the female subsample, and $\bar{y}_m$ be average of y_i in the male subsample. Let $x_i = (1, f_i)$, and let y be the $n \times 1$ vector with entries y_i, and X be the $n \times 2$ matrix with rows x_i.

(a) Somebody proposes to also include a male dummy $m_i = 1 - f_i$ into the model and to estimate the regression $y_i = \beta_1 + f_i\beta_2 + m_i\beta_3 + \epsilon_i$ by OLS. Explain why this is problematic.

(b) Somebody proposes to also include the square f_i^2 of the female dummy into the model and the estimate the regression$y_i = \beta_1 + f_i\beta_2 + f_i^2\beta_3 + \epsilon_i$ by OLS. Explain why this is problematic.

(c) Give expressions for the 2×2 matrix $X'X$ and the 2×1 vector $X'y$ in terms of n_f , n_m, y_f and y_m.

$$X'X = \begin{pmatrix} n & n_f \\ n_f & n_f \end{pmatrix} \quad X'y = \begin{pmatrix} n_f\bar{y}_f + n_m\bar{y}_m \\ n_f\bar{y}_f \end{pmatrix}$$

(d) Now assume that $n_f = n_m = 100$, that $\bar{y}_f = 165$, and that $\bar{y}_m = 175$. Calculate the OLS estimator for β_1 and β_2.

$$\beta_1 = 175 \quad \beta_2 = -10$$

(e) In addition to the assumptions in (d) now also assume that $Var(\epsilon_i|f_i) = 50$. Calculate the estimated standard error for $\hat{\beta}_2$.

The estimated standard error of $\hat{\beta}_2$ is 1

(f) Consider the null hypothesis $H_0 : \beta_2 = 0$. Using your results in (d) and (e) calculate the t-test statistics for testing H_0. Would you reject H_0 at 5% significance level?

Yes.

Problem 2

Hint: Use property CE.5.

Problem 3

(a) Load the MROZ.csv data into R (can be found on blackboard)

```
> #Loading dataset MROZ.CSV into memory
>  mroz <- read.table("MROZ.csv", header=TRUE, sep=",", na.string=".")
>  # attaches variable names to dataset, allowing for calling variables individually
>  #Be careful you do not override previous variables with the same name
>  attach(mroz)
```

(b) Run some summary statistics. Are there any variables with missing values? Why might there be missing values in this (these) variables?

```
>  summary(mroz)
```

The dataset contains missing values in the wage and lwage variables. This occurs, since the wage is observed only for those individuals who were employed at the time of the data collection (inlf=1)

(c) Estimate the following model[1]

$$\log(wage) = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 educ + \beta_4 age + \beta_5 kidslt6 + \beta_6 kidsge6 + u$$

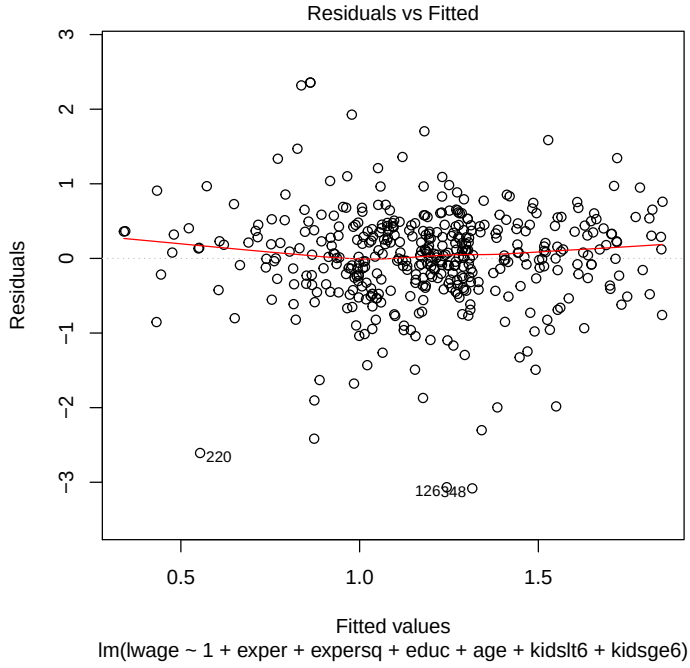
with the normal standard errors, for the 428 employed women in the sample. Compare your results with Example 4.1 in Graduate Wooldridge (2002, 2010).

```
>  mroz.lm <- lm(lwage ~ 1+exper+expersq+educ+age+kidslt6+kidsge6, data=mroz, x=T)
>  # Displaying model estimation
>  summary(mroz.lm)

Call:
lm(formula = lwage ~ 1 + exper + expersq + educ + age + kidslt6 +
    kidsge6, data = mroz, x = T)

Residuals:
     Min       1Q   Median       3Q      Max
-3.08183 -0.30631  0.04606  0.37161  2.35708

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4209080  0.3169050  -1.328  0.18484
exper        0.0398190  0.0133930   2.973  0.00312 **
expersq     -0.0007812  0.0004022  -1.942  0.05276 .
```

[1] Equation (4.16) in Wooldrige (2010,2002)

```
educ         0.1078320  0.0144021   7.487 4.16e-13 ***
age         -0.0014653  0.0052925  -0.277  0.78203
kidslt6     -0.0607106  0.0887626  -0.684  0.49437
kidsge6     -0.0145910  0.0278981  -0.523  0.60124
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.6682 on 421 degrees of freedom
  (325 observations deleted due to missingness)
Multiple R-squared:  0.1582,        Adjusted R-squared:  0.1462
F-statistic: 13.19 on 6 and 421 DF,  p-value: 1.057e-13
```

```
> # Visualization of the model (look for functional forms in the residuals)
> plot.lm(mroz.lm, which=1)
```



The red line in your plot is a "lowess smoother" – a locally weighted polynomial regerssion. A sufficient curvature is a sign of either heteroskedasticity or a model misspecification. The bptest test for heteroskedasticity.

```
> library(lmtest)
> bptest(mroz.lm)

        studentized Breusch-Pagan test
```

```
data:  mroz.lm
BP = 15.7291, df = 6, p-value = 0.01528
```

(d) Estimate the model in problem 3 with heteroscedasticity robust standard errors. Compare your results with Example 4.1.

vcovHC is a function in the package sandwich which calculates the robust standard errors of a linear model. Several types of robust standard errors can be specified. HC0 are the White's standard errors but there are other options. Type ? vcovHC (after loading the package sandwich) for more details.

```
> library(sandwich)
> coeftest(mroz.lm, vcov=vcovHC(mroz.lm, type="HC0"))

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -0.42090796  0.31572069 -1.3332  0.183198
exper        0.03981902  0.01513251  2.6314  0.008817 **
expersq     -0.00078123  0.00040632 -1.9227  0.055193 .
educ         0.10783196  0.01351167  7.9807 1.396e-14 ***
age         -0.00146526  0.00588632 -0.2489  0.803539
kidslt6     -0.06071057  0.10522938 -0.5769  0.564291
kidsge6     -0.01459101  0.02910954 -0.5012  0.616461
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

> # As an alternative for HC0, the following command can be used. This requires package
> # 'survival'
> library(survival)
> coeftest(mroz.lm, vcov=sandwich)

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -0.42090796  0.31572069 -1.3332  0.183198
exper        0.03981902  0.01513251  2.6314  0.008817 **
expersq     -0.00078123  0.00040632 -1.9227  0.055193 .
educ         0.10783196  0.01351167  7.9807 1.396e-14 ***
age         -0.00146526  0.00588632 -0.2489  0.803539
kidslt6     -0.06071057  0.10522938 -0.5769  0.564291
kidsge6     -0.01459101  0.02910954 -0.5012  0.616461
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

(e) Write down the hypothesis that education does not affect wages in both cases, and use the t-values from the above regressions to reach a conclusion about the significance of education.

$H0 : \beta_{educ} = 0, H_1 : \beta_{educ} \neq 0$. Using the t-statistics from both the standard OLS model (t=7.4872) and the model with Huber-White standard errors (t=7.9807). We reject the null-hypothesis, and conclude that education in both model specifications significantly affect wage earnings

(f) Use the F-statistics to test the hypothesis of $\beta_4 = \beta_5 = \beta_6 = 0$ as in the example.

(f.1) Run and save the restricted regression

```
>  mroz.restricted <- lm(lwage ~ 1+exper+expersq+educ, data=mroz)
```

(f.2) Use the 'anova' command to perform the F-test: "anova(*restricted_model, unrestricted_model*)"

We can also use the waldtestlmtest and the linearHypothesiscar to perform the F-test.

```
> anova(mroz.restricted, mroz.lm)

Analysis of Variance Table

Model 1: lwage ~ 1 + exper + expersq + educ
Model 2: lwage ~ 1 + exper + expersq + educ + age + kidslt6 + kidsge6
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    424 188.31
2    421 187.99  3   0.31751 0.237 0.8705

> library(lmtest)
> waldtest(mroz.restricted, mroz.lm)

Wald test

Model 1: lwage ~ 1 + exper + expersq + educ
Model 2: lwage ~ 1 + exper + expersq + educ + age + kidslt6 + kidsge6
  Res.Df Df     F Pr(>F)
1    424
2    421  3 0.237 0.8705

> library(car)
> linearHypothesis(mroz.lm, c("age=0", "kidslt6=0", "kidsge6=0"))

Linear hypothesis test

Hypothesis:
age = 0
kidslt6 = 0
kidsge6 = 0

Model 1: restricted model
Model 2: lwage ~ 1 + exper + expersq + educ + age + kidslt6 + kidsge6
```

```
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    424 188.31
2    421 187.99  3   0.31751 0.237 0.8705
> # As a bonus, the waldtest and linearHypotehesis commands
> #is that they allow robust F-tests
> waldtest(mroz.restricted, mroz.lm,
+          vcov=vcovHC(mroz.lm, type="HC0"))
Wald test

Model 1: lwage ~ 1 + exper + expersq + educ
Model 2: lwage ~ 1 + exper + expersq + educ + age + kidslt6 + kidsge6
  Res.Df Df      F Pr(>F)
1    424
2    421  3 0.1672 0.9185
> linearHypothesis(mroz.lm, c("age=0", "kidslt6=0", "kidsge6=0"),
+                  vcov=vcovHC(mroz.lm, type="HC0"))
Linear hypothesis test

Hypothesis:
age = 0
kidslt6 = 0
kidsge6 = 0

Model 1: restricted model
Model 2: lwage ~ 1 + exper + expersq + educ + age + kidslt6 + kidsge6

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1    424
2    421  3 0.1672 0.9185
```

The p-value of the F-test is 0.8705 we cannot reject the null-hypothesis, i.e. we cannot reject that the coefficients on age, kidsle6 and kidsge6 can all be equal to zero

(g) Test the same hypothesis using the LM statistic

(g.1) Extract the residuals from the restricted regression

```
>  mroz.residual= resid(mroz.restricted)
>  length(mroz.residual)
[1] 428
>  nrow(mroz)
[1] 753
```

```
>
```

(g.2) Regress the restricted residuals on the full set of explanatory variables (including the variables you are testing)

The length of the database and the residuals is different becuase when running the lm model, this has removed the missing values. That is why the command below

We tell R to run the regression with the dataset without missing values:

```
> lmtest.lm <- lm(mroz.residual ~ 1+exper+expersq+educ+age+
+                                  kidslt6+kidsge6, data=na.exclude(mroz))
> summary(lmtest.lm)

Call:
lm(formula = mroz.residual ~ 1 + exper + expersq + educ + age +
    kidslt6 + kidsge6, data = na.exclude(mroz))

Residuals:
     Min       1Q   Median       3Q      Max
-3.08183 -0.30631  0.04606  0.37161  2.35708

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.011e-01  3.169e-01   0.319    0.750
exper       -1.747e-03  1.339e-02  -0.130    0.896
expersq      2.996e-05  4.022e-04   0.074    0.941
educ         3.423e-04  1.440e-02   0.024    0.981
age         -1.465e-03  5.292e-03  -0.277    0.782
kidslt6     -6.071e-02  8.876e-02  -0.684    0.494
kidsge6     -1.459e-02  2.790e-02  -0.523    0.601

Residual standard error: 0.6682 on 421 degrees of freedom
Multiple R-squared:  0.001686,        Adjusted R-squared:  -0.01254
F-statistic: 0.1185 on 6 and 421 DF,  p-value: 0.9942
```

(g.3) The test statistic is then $N * R^2$, where N is the number of observations and R^2 is the R squared from the regression from step b. Under the null, this is chi-squared distributed with k degrees of freedom, where k is the number of restrictions (in this case k=3).[2]

```
>  r.squared<-summary(lmtest.lm)$r.squared
>  lmtest.statistic <- 428*r.squared
>  pchisq(lmtest.statistic, 3, lower.tail=F)

[1] 0.8680941
```

[2]Use the *pchisq(X, df=k)* command to display the cumulative chi-squared distribution function where X is the test statistic and k is the degree of freedom. The probability of the null hypothesis is then *1-pchisq(X, df=k)*.

A heteroskedastic robust LM test(see page 61-64 of Wooldridge (2010))

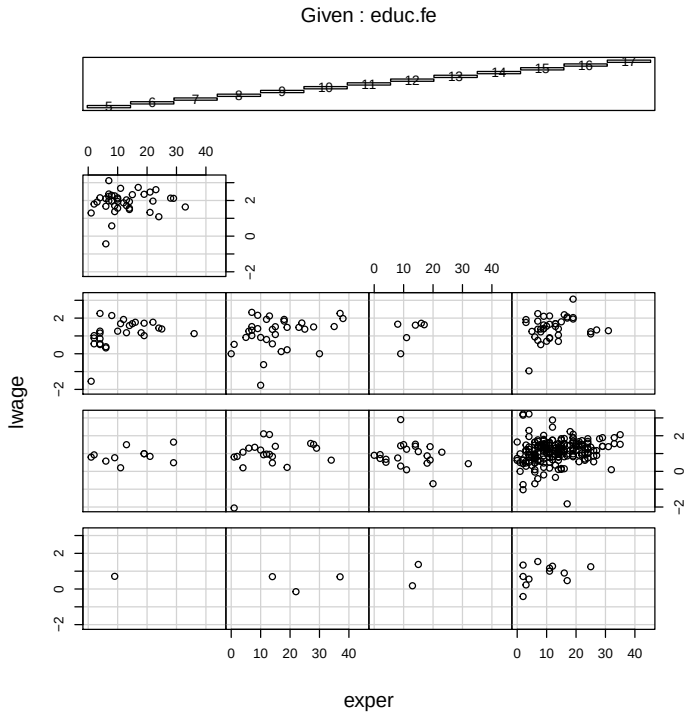
```
> # Step a: extracting residuals from restricted regression
>         mroz.restricted <- lm(lwage ~ 1+exper+expersq+educ, data=mroz)
>   # Creating dataset of non-missing observations
>         mroz.nona <- data.frame(mroz.nona, mroz.residual = resid(mroz.restricted))
> # Step b: Regress each of the restricted variables on the included explanatory
>   # variables and get residuals from these
>         age.lm <- lm(age ~ 1+exper+expersq+educ, data=mroz.nona)
>         mroz.nona <- data.frame(mroz.nona, res.age = residuals(age.lm))
>         kidslt6.lm <- lm(kidslt6 ~ 1+exper+expersq+educ, data=mroz.nona)
>         mroz.nona <- data.frame(mroz.nona, res.kidslt6 = residuals(kidslt6.lm))
>         kidsge6.lm <- lm(kidsge6 ~ 1+exper+expersq+educ, data=mroz.nona)
>         mroz.nona <- data.frame(mroz.nona, res.kidsge6 = residuals(kidsge6.lm))
> # Step c: Generate new variables being the interaction between the restricted
> # residuals and the residuals from the three regressions above
>         attach(mroz.nona)
>         mroz.nona <- data.frame(mroz.nona, var1 = mroz.residual*res.age)
>         mroz.nona <- data.frame(mroz.nona, var2 = mroz.residual*res.kidslt6)
>         mroz.nona <- data.frame(mroz.nona, var3 = mroz.residual*res.kidsge6)
> # Step d: regress a unit vector on these three variables with no constant
>         mroz.nona <- data.frame(mroz.nona, unit = 1)
>         het.lm <- lm(unit~0+var1+var2+var3, data=mroz.nona)
>         # The test statistic is N-SSR from this regression, which under the null is
>   # chisquared distributed with k degrees of freedom
>   # To get the SSR (sum of squared residuals) use the 'anova()' command
>         anova(het.lm)
>         het.lm.test <- 428-427.49
>         pchisq(het.lm.test,3, lower.tail=F)
>         # Which gives us a p=0.916689, in other words, we once again cannot reject the
>   # null hypothesis
```

(h) Plot wage-experience profiles for different education levels. Interpret them.
Simple graphics of wage-experience distributions

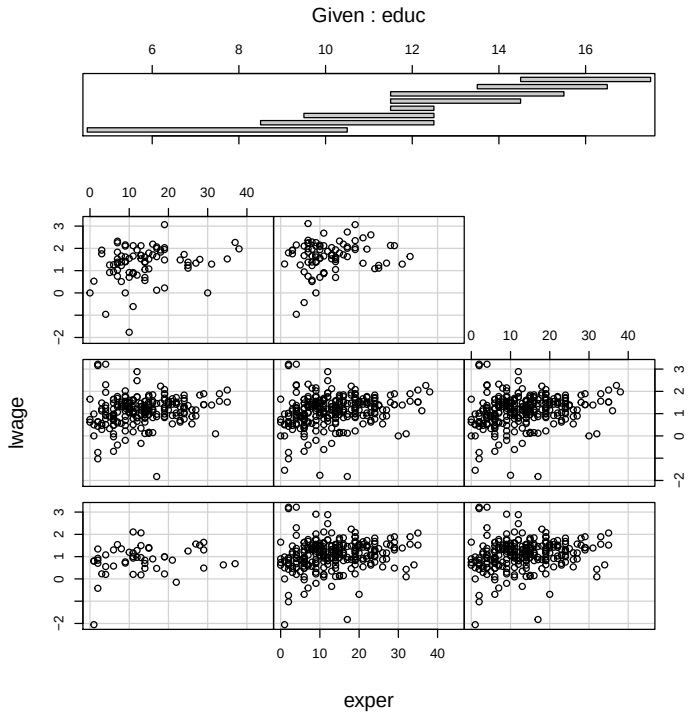
```
> mroz.nona=na.exclude(mroz)

> plot(lwage~exper, data=mroz.nona)
```

Graphics for each education level

```
>  educ.fe <- as.factor(educ)
>  num <- length(unique(educ))
>  coplot(lwage~exper | educ.fe, num=num)

 Missing rows: 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 44
```




```
> coplot(lwage~exper | educ, num=num)

 Missing rows: 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 44
```



It seems that wage increases with experience but not that much. It is definitely not clear that the nature of the relationship between time and experience depends on the level of education. This mean that if one were to fit a simple regression to each panel of the coplot, would the slopes be identical in every case?

(i) Test the hypothesis of no effect of experience on wages (note that there is both an '*exper*' and '*expersq*' term in the regression).

```
> mroz.res2  <- lm(lwage~1+educ+age+kidslt6+kidsge6, data=mroz)
>  mroz.unres <- lm(lwage~1+exper+expersq+educ+age+kidslt6+kidsge6, data=mroz)
>  anova(mroz.res2, mroz.unres)

Analysis of Variance Table

Model 1: lwage ~ 1 + educ + age + kidslt6 + kidsge6
Model 2: lwage ~ 1 + exper + expersq + educ + age + kidslt6 + kidsge6
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    423 194.45
2    421 187.99  2    6.4648 7.239 0.0008109 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

The p-value is 0.000811, indicating that we cannot reject the null – hence experience is significant in the model

(j) Include an interaction term between education and experience. How would you interpret that? Is it significant?

```
> educexper <- educ*exper
> mroz.lm2 <- lm(lwage~1+exper+educ+educexper+age+kidslt6+kidsge6)
> summary(mroz.lm2)

Call:
lm(formula = lwage ~ 1 + exper + educ + educexper + age + kidslt6 +
    kidsge6)

Residuals:
     Min       1Q   Median       3Q      Max
-3.04453 -0.30960  0.05473  0.39587  2.31063

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1029288  0.4264084  -0.241 0.809374
exper        0.0070868  0.0223833   0.317 0.751696
educ         0.1006785  0.0273595   3.680 0.000264 ***
educexper    0.0006741  0.0017560   0.384 0.701260
age         -0.0035540  0.0052176  -0.681 0.496143
kidslt6     -0.0733498  0.0889922  -0.824 0.410277
kidsge6     -0.0187925  0.0281279  -0.668 0.504429
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.6711 on 421 degrees of freedom
  (325 observations deleted due to missingness)
Multiple R-squared:  0.151,        Adjusted R-squared:  0.1389
F-statistic: 12.48 on 6 and 421 DF,  p-value: 5.863e-13
```

(k) Can you make a model that fits better than the one in problem 3?

```
> model.lm.hours<-lm(lwage~exper+expersq+educ+hours+age+kidslt6+kidsge6, x=T)
> summary(model.lm.hours)
> model.lm.hushrs<-lm(lwage~exper+expersq+educ+hushrs+age+kidslt6+kidsge6, x=T)
> summary(model.lm.hushrs)
> model.lm.faminc<-lm(lwage~exper+expersq+educ+faminc+age+kidslt6+kidsge6, x=T)
> summary(model.lm.faminc)
> model.lm.city<-lm(lwage~exper+expersq+educ+city+age+kidslt6+kidsge6, x=T)
> summary(model.lm.city)
```

(l) Discuss whether assumptions OLS.1 - OLS.3 are likely to be fulfilled. OLS1: Is the error uncorrelated with all explanatory variables? No, education might be well related to the ability of the person, the same with the number of kids and the race of the person. OLS2: Perfect collenarity? Yes OLS3: Homoscedasticity: $E(\epsilon^2 X'X) = \sigma^2$ (constant). Check plot