Econometrics exam project

Johan Bysted, Jonathan Arve & Mathias Kold

14. June 2024

Contents

Exam 1 - OLS and heteroskedasticity

In a multiple linear regression (MLR) there are 6 assumptions. Assumption 1-5 are called Gauss-Markov assumptions and assumption 6 is called the normality assumption. The first 4 assumptions exist to secure that the model is unbiased, while assumption 5 checks for heteroskedasticity and assumption 6 checks for normality in the model. The assumptions are:

MLR1: Linear in Parameters

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters of interest and u is an unobserved random error or disturbance term.

MLR2: Random Sampling

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

MLR3: No Perfect Collinearity

In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables.

MLR4: Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0$$

MLR5: Homoskedasticity

The error u has the same variance given any value of the explanatory variables. In other words,

$$\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$$

MLR6: Normality

The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0, \sigma^2)$.

Look at the following model for bank employees wage:

$$log(salary) = \beta_0 + \beta_1 educ + \beta_2 log(salbegin) + \beta_3 male + \beta_4 minority + u$$

where salary is yearly wage (in 1000 US dollars), educ is education measured in number of years, salbegin is the starting salary (in 1000 US dollars) for the person's first position in the same bank, male is a dummy variable for gender, minority is one dummy variable indicating whether one belongs to a minority.

1.1 - Estimate the model using OLS. Comment on the output and interpret the results

Here we use the lm() function to estimate the model.

```
model <- lm(log(salary) ~ educ + log(salbegin) + male + minority, data = data1)
summary(model)
```

```
##
## Call:
## lm(formula = log(salary) ~ educ + log(salbegin) + male + minority,
##     data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45572 -0.11508 -0.00516  0.10765  0.87060
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    0.84868    0.07512  11.298 < 0.0000000000000002 ***
## educ           0.02327    0.00387   6.013        0.00000000366 ***
## log(salbegin)  0.82180    0.03603  22.808 < 0.0000000000000002 ***
## male           0.04816    0.01991   2.419               0.0160 *
## minority      -0.04237    0.02034  -2.083               0.0378 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1766 on 469 degrees of freedom
## Multiple R-squared:  0.8041, Adjusted R-squared:  0.8024
## F-statistic: 481.3 on 4 and 469 DF,  p-value: < 0.00000000000000022
```

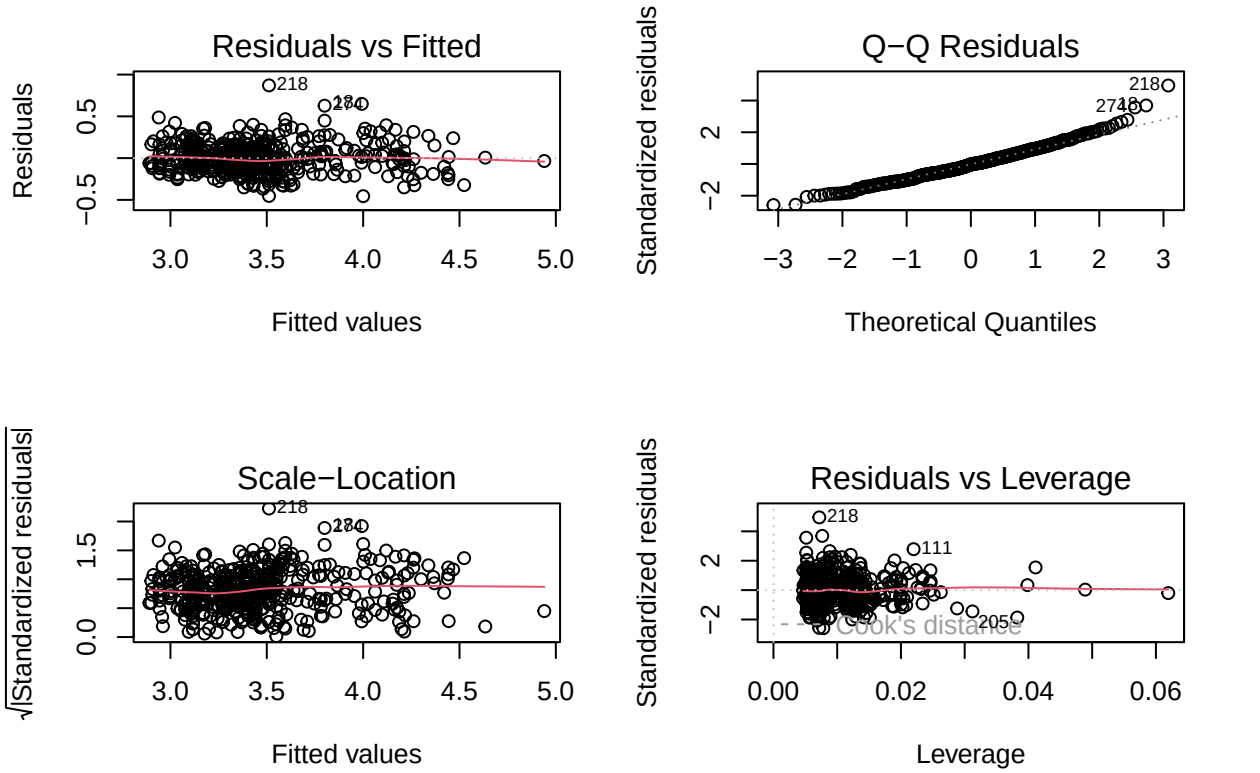
From the estimated model it can be seen that the education variable has a value of 0.02327 which means that one extra year of education will raise the salary by approximately 2.3%. The male variable has a value 0.04816 which means that being a man will raise your salary by approximately 4.8%. The minority variable has a value of -0.04237 which means that minorities approximately will have a 4.2% lower salary than people who are not minorities. The salbegin variable has a different interpretation because it is in log form. It has a value of 0.82180 which means that an increase of 1% in your starting salary will make your salary approximately 0.82% higher.

The intercept is 0.84868 which is the value of the dependent variable, in this case salary, when all other variables have a value of 0. So in this case 0.84868 is the expected value of the salary for a person with zero education, without a starting salary and someone who is not a male or a minority.

It can also be seen that all the variables are statistically significant at a 5% significance level due to p-values < 0.05 but only education and log(salbegin) are statistically significant at a 1% significance level.

1.2 - Perform graphical model checking.

```
par(mfrow = c(2, 2))
plot(model)
```



The residual vs fitted model shows if the residuals have non-linear patterns. If they are equally spread around a horizontal line without any patterns it indicates that the residuals does not have non-linear patterns. From the plot it can be seen that the spread is quite equal around the horizontal red line, although a few outliers exists, which indicates a linear pattern.

The Q-Q residuals plot shows if the residuals are normally distributed. If the residuals fit the dotted line they are normally distributed. In this case, the residuals fit the dotted line except in the last quantiles where they deviate a bit from the dotted line. This could indicate that the residuals are approximately normally distributed although not perfectly normally distributed.

The scale-location plot shows if homoskedasticity exists within the residuals. If the residuals are spread equally around the predictors, the model fulfills the assumption of homoskedasticity. In this case, it can be argued that there is an increasing pattern in the spread of the residuals which indicates heteroskedasticity.

The residuals vs leverage plot helps detect outliers in the model. It is important because outliers can have a major impact in the model. The residuals can be seen if they have a large Cook's distance, which is the dashed line. In this case, it seems like there might be an outlier with observation 218 but other than that, most of the observations does not have a large Cook's distance.

1.3 - Test for heteroskedasticity using the Breusch-Pagan test and the special edition of the White test.

When testing for heteroskedasticity, a hypothesis can be made from the MLR5 assumption about homoskedasticity in the model:

$$\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$$

Given the MLR4 assumption that the error u has a expected value of zero given any values of the independent variables, it can be rewritten as:

$$\text{Var}(u) = E\left[(u - E(u))^2\right] = E\left[u^2\right] - \left(E\left[u\right]\right)^2$$

When the expected value of u is zero it can be rewritten as:

$$E\left(u^2 \mid x_1, \ldots, x_k\right) = \sigma^2$$

To test whether this applies, a hypothesis test can be performed:

$$H_0 : E\left(u^2 \mid x_1, \ldots, x_k\right) = \sigma^2$$

$$H_1 : E\left(u^2 \mid x_1, \ldots, x_k\right) \neq \sigma^2$$

This is tested by examining whether some of the independent variables are correlated with the u^2. For this we set up a regression model. Here it must be noted that the squared error term of the population is not known, why this is estimated from the sample with the following method:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + error$$

On the regression is then performed a F-test/LM test, to test whether the coefficients are equal to 0. If they are, then the null hypothesis can not be rejected, indicating homoskedasticity. If they are not equal to 0, then the alternative hypothesis is accepted indicating heteroskedasticity in the model.

When we want to test multiple hypothesis we turn to the F-statistics, where we use the following formula:

$$F = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} * \frac{n - k - 1}{q}$$

where R_{ur}^2 is the R-squared from our unrestricted model, R_r^2 is the R-squared from the restricted model, q is the difference in the degrees of freedom between the unrestricted and restricted model, n is the number of observations in the dataset and k is the number of independent variables in the unrestricted model.

We perform this directly in R with the summary():

```
r=residuals(model)
res = r^2
summary(lm(res~educ + log(salbegin) + male + minority, data = data1))
```

```
##
## Call:
## lm(formula = res ~ educ + log(salbegin) + male + minority, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04773 -0.02506 -0.01345  0.00908  0.71750
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.0036404  0.0235992   0.154    0.877
## educ           0.0019699  0.0012157   1.620    0.106
## log(salbegin) -0.0008061  0.0113201  -0.071    0.943
## male           0.0094827  0.0062553   1.516    0.130
## minority      -0.0104497  0.0063908  -1.635    0.103
##
## Residual standard error: 0.05548 on 469 degrees of freedom
## Multiple R-squared:  0.02923,    Adjusted R-squared:  0.02095
## F-statistic: 3.531 on 4 and 469 DF,  p-value: 0.007475
```

Because the p-value is 0.007475 which is < 0.05, we reject the null hypothesis meaning that there is heteroskedasticity in the model.

The LM test is simply obtained by: $LM = R^2_{ur} * n$ We can now calculate the LM test, where we have 474 observations and $R^2 = 0.02923$:

```
LM = 0.02923*474
LM
```

```
## [1] 13.85502
```

We can then calculate the p-value of chi-square χ^2_k from LM test in R:

```
1-pchisq(LM,4)
```

```
## [1] 0.007772441
```

Both the LM and F-test reject the null hypothesis meaning there is heteroskedasticity in the model. The Breusch-Pagan test can also be directly run in R using the bptest function which also gives us the p-value.

```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 13.857, df = 4, p-value = 0.007767
```

The special edition of the White test The special White test can also be used to do the test the model for heteroskedasticity. It is a simplification of the original proposed test that adds the squares and cross products of all the independent variables to the equation. This would become more and more complicated as the number of independent variables in the model increase, so instead the special edition of the White test is more simple and maintains a low number of degrees of freedom. It tests whether the error term is correlated with some of the included variables, which will create heteroskedasticity. Practically, adapted values of y are included in this test:

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y} + e$$

Then we test the joint hypothesis: $H_0 : \delta_1 = 0, \delta_2 = 0$ against $H_1 : \delta_1 \neq 0, \delta_2 \neq 0$. If the null hypothesis is accepted, then homoscedasticity is assumed, as the result shows that the fitted values of the model is not correlated with the error term, which indicates that there is no heteroskedasticity:

```
yhat <- fitted(model)
quadu <- (residuals(model)^2)
model_white <- lm(quadu ~ yhat + I(yhat^2))

whitetest <- c("yhat=0", "I(yhat^2)=0")
linearHypothesis(model_white,whitetest)
```

```
## Linear hypothesis test
##
## Hypothesis:
## yhat = 0
## I(yhat^2) = 0
##
## Model 1: restricted model
## Model 2: quadu ~ yhat + I(yhat^2)
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    473 1.4873
## 2    471 1.4531  2  0.034139 5.5327 0.004217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From both the BP test and the special White test it can be seen that the p-value is < 0.05 which means that there is heteroskedasticity in the model.

1.4 - Calculate robust standard errors for the model and compare with the results in question 1.

Since there is heteroskedasticity in the model, it is necessary to take this into account, when performing a linear regression. A valid estimator of multiple linear regression in the presence of heteroskedasticity is

$$Var(\hat{\beta}_j) = \frac{\sum_{i=0}^{n} \hat{r}_{ij}^2 \cdot \hat{u}_i^2}{(SSR_j)^2}$$

where $\hat{r}_{ij}$ denotes the ith residual from regressing x_j on all other independent variables. The robust standard error is obtained by taking the square root of the above equation. R can calculate the robust standard errors using coeftest():

```
coef_model <- coeftest(model, vcov=vcovHC(model,type="HC0"))
screenreg(list(OLS=model,Standard_Robust_Error=coef_model), digits=4)
```

```
##
## ================================================
##                OLS           Standard_Robust_Error
## ------------------------------------------------
## (Intercept)     0.8487 ***    0.8487 ***
##                (0.0751)      (0.0794)
## educ            0.0233 ***    0.0233 ***
##                (0.0039)      (0.0035)
## log(salbegin)   0.8218 ***    0.8218 ***
##                (0.0360)      (0.0374)
## male            0.0482 *      0.0482 *
##                (0.0199)      (0.0200)
## minority       -0.0424 *     -0.0424 *
##                (0.0203)      (0.0177)
## ------------------------------------------------
## R^2             0.8041
## Adj. R^2        0.8024
## Num. obs.      474
## ================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

It can be seen that the estimates from the two models are the same, which is expected, because we try to account for the non-constant variance of the error term in the OLS-model. Hence, it is only the standard errors that change. Furthermore, the significance level does not change, when performing a coeftest(), when adjusting for heteroskedasticity. The robust standard errors can be used to perform hypothesis testing and to calculate confidence intervals. Even when adjusting for heteroskedasticity, the conclusion from the regressions does not change, since the estimates are the same, and there are no significant changes.

1.5 - Test the hypothesis $H_0 : \beta_2 = 1$ against the alternative $H_1 : \beta_2 \neq 1$.

β_2 is the estimate for the log-starting salary's effect on the the yearly salary, so when we want to test the null hypothesis $H_0 : \beta_2 = 1$ it means that we want to test whether the log-starting salary population parameter is equal to 1 (indicating a 1% increase in log-starting salary will make log-salary 1% higher).

To test our null hypothesis against the alternative hypothesis we use the t-statistics, which is given by:

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - 1}{se(\hat{\beta}_j)}$$

We will perform a two-sided test, so the decision rule will be $|t_{\hat{\beta}_j}| > c$, meaning that the null hypothesis will be rejected if the absolute value of t-statistics is greater than the critical value.

To perform the test we use our estimate and standard error for $\hat{\beta}_2$ from 1.1.

```
bhat2 <- 0.82180
se_bhat2 <- 0.03603

t_stat <- (bhat2 - 1) / se_bhat2
t_stat
```

```
## [1] -4.945878
```

Then we need to calculate the critical values for the two-sided test:

```
alpha <- 0.05
c <- qt(1-alpha/2, 469)
c
```

```
## [1] 1.965035
```

```
abs(t_stat)>c
```

```
## [1] TRUE
```

We can see that the absolute value of the t-statistic is greater than the critical value at a 5% significance level, meaning that we reject the null hypothesis, so we instead accept the alternative hypothesis saying that the true population parameter for log-starting salary is not 1.

Another way to test the null hypothesis is by calculating the p-value. The definition of the p-value is the probability of obtaining a t-statistic more or as extreme as the one observed in the sample. We use R to calculate the p-value in the following way:

```
pval <- 2*pt(-abs(t_stat), 469)
pval
```

```
## [1] 0.000001057867
```

So at a 5% significance level the p-value also tells us to reject the null hypothesis, further confirming the rejecting from the t-statistics before.

1.6 - Test the hypothesis $H_0 : \beta_3 = \beta_4 = 0$

When we want to test multiple hypothesis we turn to the F-statistics, where we use the following formula:

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} * \frac{n - k - 1}{q}$$

where R^2_{ur} is the R-squared from our unrestricted model, R^2_r is the R-squared from the restricted model, q is the difference in the degrees of freedom between the unrestricted and restricted model, n is the number of observations in the dataset and k is the number of independent variables in the unrestricted model.

In our case the unrestricted model is the one given in the beginning of assignment 1.1 with all the independent variables, and in the case where we want to test the null hypothesis $H_0 : \beta_3 = \beta_4 = 0$, we then get the following restricted model:

$$log(salary) = \beta_0 + \beta_1 educ + \beta_2 log(salbegin) + u$$

So we start by estimating our restricted model:

```
model_r <- lm(log(salary) ~ educ + log(salbegin), data = data1)
```

We then obtain our R^2 from the restricted and unrestricted model:

```
r2_ur <- summary(model)$r.squared
r2_r <- summary(model_r)$r.squared
```

We now have what we need to calculate our F-statistic:

```
F <- (r2_ur-r2_r)/(1-r2_ur) * (474-4-1)/2
F
```

```
## [1] 4.234946
```

We then calculate our critical values of the F-statistics at 5% significance level:

```
qf(0.95, 2, 469)
```

```
## [1] 3.014949
```

Our decision rule says that if F > c then we reject our null hypothesis and instead accept our alternative hypothesis meaning that β_3 and β_4 does have a statistical impact on yearly salary.

It is also possible to do the test directly in R by using the following code:

```
myh0 <- c("male=0", "minority=0")
linearHypothesis(model, myh0)
```

```
## Linear hypothesis test
##
## Hypothesis:
## male = 0
## minority = 0
##
## Model 1: restricted model
## Model 2: log(salary) ~ educ + log(salbegin) + male + minority
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    471 14.892
## 2    469 14.627  2   0.26416 4.2349 0.01504 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the code in R to perform the F-statistics gives us the same answer as before still confirming the rejecting of our null hypothesis.

1.7 - Estimate the model using FGLS and comment on the results.

Feasible generalised least squares (FGLS) is a method used to adress heteroskedasticity. It can be difficult to know the form of heteroskedasticity (i.e, h(x_i)). In many cases, it is possible to model the function h and use this data to estimate the unknown parameters. By estimating h_i, an estimated value of h is obtained, $\hat{h}_i$. One method to find the exact form of h_i is as follows:

$$Var(u|x) = \sigma^2 exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k)$$

where $x_1, x_2, ..., x_k$ are the independent variables in the regression.

Since the parameters δ is unknown for the population, these are estimated using the given data, where $\hat{h}$ can be used to account for heteroskedasticity. To find the estimates for δ, we can use the following:

$$u^2 = \sigma^2 exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k)v$$

By taking log, the model can be linearised:

$$log(u^2) = a_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + e$$

We already know from 1.3 that our model has heteroskedasticity problems, so we obtain our squared residuals from our OLS model and log them:

```
logu2 <- log(resid(model)^2)
```

Now we run the regression on the form mentioned earlier:

```
varreg <- lm(logu2 ~ educ + log(salbegin) + male + minority, data = data1)
```

After we calculate the weights by exponentiating the fitted values from varreg:

```
w <- exp(fitted(varreg))
```

And then we can estimate the FGLS:

```
FGLS <- lm(log(salary) ~ educ + log(salbegin) + male + minority, weight=1/w, data = data1)
screenreg(list(OLS=model, FLGS=FGLS), digits=4)
```

```
##
## ========================================
##                OLS           FLGS
## ----------------------------------------
## (Intercept)     0.8487 ***    0.8493 ***
##                (0.0751)      (0.0756)
## educ            0.0233 ***    0.0222 ***
##                (0.0039)      (0.0038)
## log(salbegin)   0.8218 ***    0.8270 ***
##                (0.0360)      (0.0358)
## male            0.0482 *      0.0487 *
##                (0.0199)      (0.0196)
## minority       -0.0424 *     -0.0429 *
##                (0.0203)      (0.0187)
## ----------------------------------------
## R^2             0.8041        0.8046
## Adj. R^2        0.8024        0.8029
## Num. obs.      474           474
## ========================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

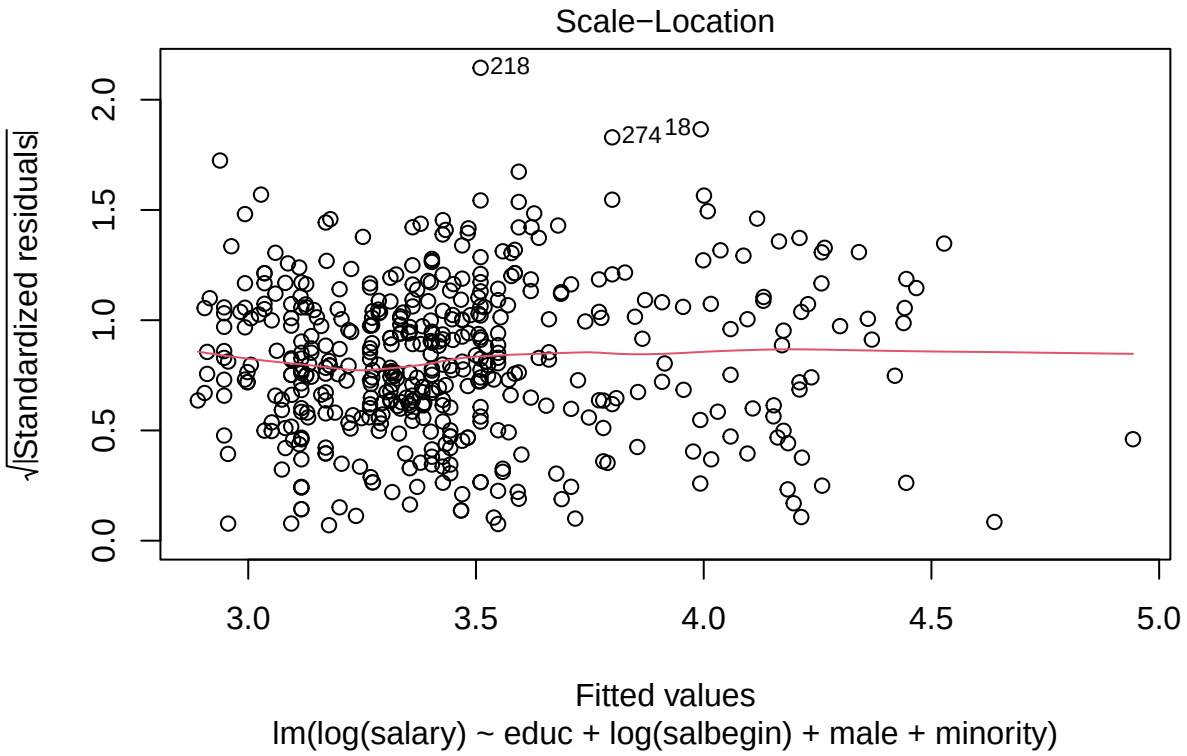
It can be seen that the standard errors for the different variables have been changed a bit, but nothing significant. It is very small changes, which maybe could indicate that the FGLS estimation have not taken all the heteroskedasticity into account. This will be elaborated further in 1.8.

1.8 - Has the FGLS estimation taken into account all the heteroskedasticity?

There are more ways to check whether the FGLS estimation has taken all the heteroskedasticity into account. One way is to look at it graphically and another is by doing a BP-test.

We start by looking at it graphically in the Scale-Location plot:

```
plot(FGLS, 3)
```



Then we perform a BP-test:

```
bptest(FGLS)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  FGLS
## BP = 60112, df = 4, p-value < 0.00000000000000022
```

Both the plot and BP-test indicates that the FGLS estimation have not taken all of the heteroskedasticity into account. In the plot it looks like there is a slight upward trend and that the spread of the residuals increase with higher fitted values. This suggest that the variance of the residuals is not constant indicating heteroskedasticity. As mentioned in 1.3 the null hypothesis in a BP-test is homoskedasticity, but we reject the null hypothesis here, since p-value is < 0.05.

Exam 2 - OLS and misspecification

In a multiple linear regression (MLR) there are 6 assumptions. Assumption 1-5 are called Gauss-Markov assumptions and assumption 6 is called the normality assumption. The first 4 assumptions exist to secure that the model is unbiased, while assumption 5 checks for heteroskedasticity and assumption 6 checks for normality in the model. The assumptions are:

MLR1: Linear in Parameters

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters of interest and u is an unobserved random error or disturbance term.

MLR2: Random Sampling

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

MLR3: No Perfect Collinearity

In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables.

MLR4: Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0$$

MLR5: Homoskedasticity

The error u has the same variance given any value of the explanatory variables. In other words,

$$\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$$

MLR6: Normality

The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0, \sigma^2)$.

Consider the following two models for bank employees' salaries:

1. $salary = \beta_0 + \beta_1 educ + \beta_2 salbegin + \beta_3 male + \beta_4 minority + u$
2. $\log(salary) = \beta_0 + \beta_1 educ + \beta_2 \log(salbegin) + \beta_3 male + \beta_4 minority + u$

where *salary* is the annual salary (in 1000 US dollars), *educ* is education measured in number of years, *salbegin* is the starting salary (in 1000 US dollars) for the person's first position in the same bank, *male* is a dummy variable for gender, and *minority* is a dummy variable indicating whether one belongs to a minority.

```
educ <- data2$educ; salbegin <- data2$salbegin; male <- data2$male
minority <- data2$minority; salary <- data2$salary
```

2.1. Estimate the two models using OLS. Comment on the output, compare and interpret the results.

Here we use the lm() function to estimate the two models.

```
model = lm(salary ~ educ + salbegin + male + minority, data = data2)
model2 = lm(log(salary) ~ educ + log(salbegin) + male + minority, data = data2)
```

```
summary(model)
```

```
##
## Call:
## lm(formula = salary ~ educ + salbegin + male + minority, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.470  -4.128  -0.705   2.888  48.718
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -6.93228    1.85539  -3.736             0.000211 ***
## educ         0.99327    0.16674   5.957        0.00000000522 ***
## salbegin     1.60816    0.06408  25.097 < 0.0000000000000002 ***
## male         1.83088    0.85713   2.136             0.033220 *
## minority    -1.72539    0.92056  -1.874             0.061547 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.875 on 445 degrees of freedom
## Multiple R-squared:  0.7962, Adjusted R-squared:  0.7944
## F-statistic: 434.7 on 4 and 445 DF,  p-value: < 0.00000000000000022
```

The education variable has an estimate of 0.99327 which means that one extra year of education will raise the annual salary by approximately 993 dollars. The salbegin variable has an estimate of 1.60816 which means that one more unit (1000 dollars) of starting salary will give 1.60816 more units (1608.16 dollars) of annual salary. The male variable has an estimate of 1.83088 which means that if you are a male your annual salary will be 1830.88 dollars higher than if you are a woman. The minority variable has an estimate of -1.72539 which means that if you are a minority your annual salary will be 1725.39 dollars lower than if you are not a minority.

The intercept is -6.93228 which is the value of the dependent variable, in this case annual salary (in 1000 dollars), when all other variables have a value of 0. In this case a negative intercept does not really make sense since salary can not be negative.

It can also be seen that all the variables except minority variable are statistically significant at a 5% significance level due to p-values < 0.05. Education and salbegin are also significant at a 1% significance level while the male variable is not.

```
summary(model2)
```

```
##
## Call:
## lm(formula = log(salary) ~ educ + log(salbegin) + male + minority,
##     data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45488 -0.11663 -0.00496  0.11201  0.87115
##
## Coefficients:
##                Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    0.849130   0.077094  11.014 < 0.0000000000000002 ***
## educ           0.023578   0.003993   5.905        0.00000000701 ***
## log(salbegin)  0.820725   0.037051  22.151 < 0.0000000000000002 ***
## male           0.045474   0.020774   2.189               0.0291 *
## minority      -0.041856   0.021057  -1.988               0.0474 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1786 on 445 degrees of freedom
## Multiple R-squared:  0.8051, Adjusted R-squared:  0.8034
## F-statistic: 459.6 on 4 and 445 DF,  p-value: < 0.00000000000000022
```

In the second model, it can be seen that the education variable has an estimate of 0.023578 which means that one extra year of education will raise the annual salary by approximately 2.36%. The log(salbegin) variable has an estimate of 0.820725. Since it is in log form, it means that a 1% increase in the starting salary will raise the annual salary by approximately 0.821%. The male variable has an estimate of 0.045474 which means that your annual salary will be approximately 4.55% higher if you are a male than if you are a woman. The minority variable has an estimate of -0.041856 which means that if you are a minority your annual salary will be approximately 4.19% lower than if you are not a minority.

The intercept is 0.84913 which is the value of the dependent variable, in this case annual salary (in 1000 dollars), when all other variables have a value of 0.

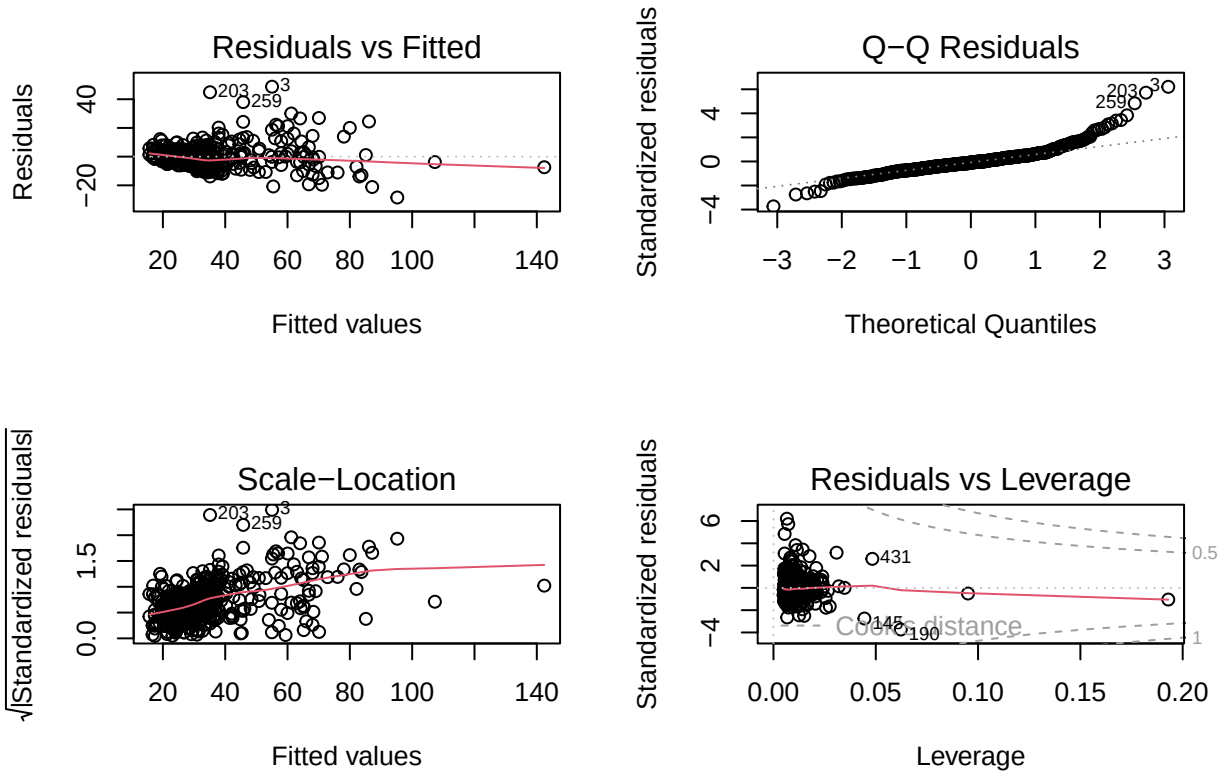
It can also be seen that all the variables are statistically significant at a 5% significance level due to p-values < 0.05. Only education and log(salbegin) are significant at a 1% significance level.

It is worth noting that the adjusted R-squared is 0.8034 for the second model, where it is a bit lower in the first model at 0.7944.

2.2. Carry out graphical model checking of the two models. Which model would you prefer?

Model 1

```
par(mfrow = c(2, 2))
plot(model)
```



The residual vs fitted model shows if the residuals have non-linear patterns. If they are equally spread around a horizontal line without any patterns it indicates that the residuals does not have non-linear patterns. It can be seen that the residuals are not quite spread around a horizontal line but rather a decreasing line.

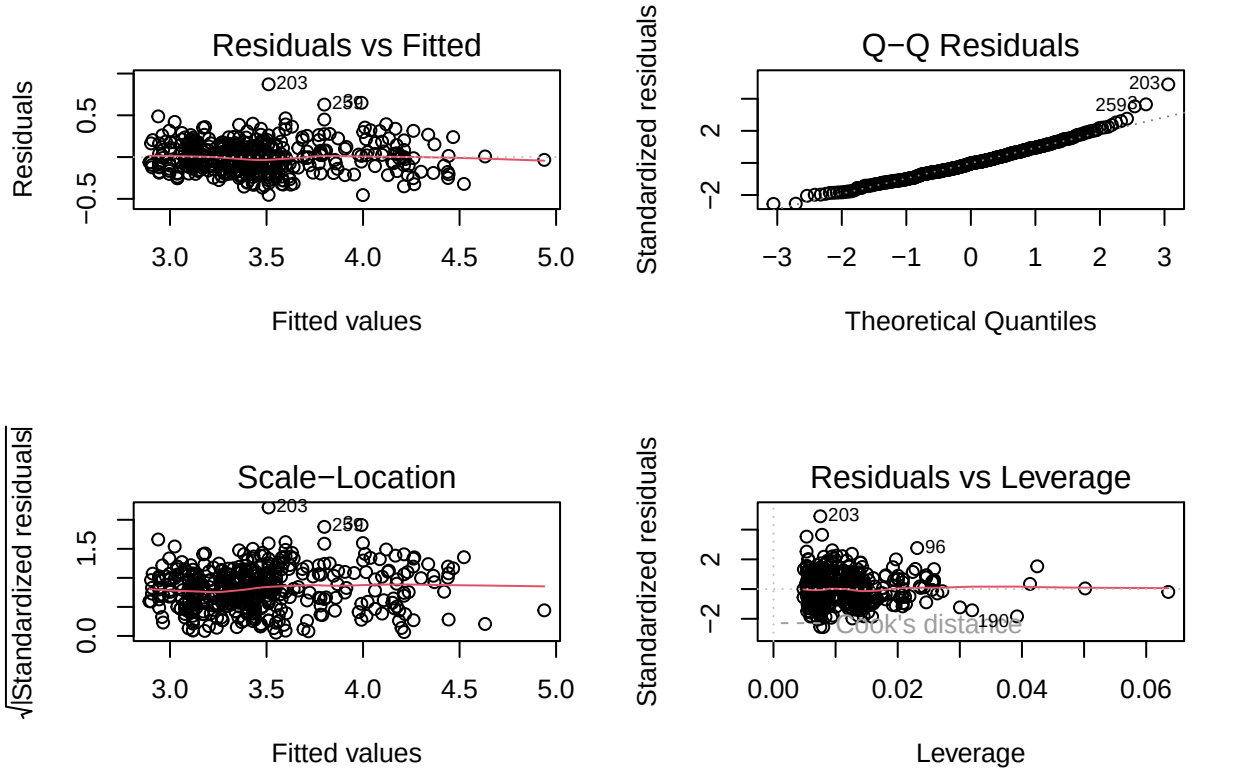
The Q-Q residuals plot shows if the residuals are normally distributed. If the residuals fit the dotted line they are normally distributed. In this case, the residuals fit the dotted line except in the first and last quantiles where they deviate from the dotted line. They deviate a bit more upwards from the line than downwards. This could indicate that the residuals are approximately normally distributed although not perfectly normally distributed.

The scale-location plot shows if homoskedasticity exists within the residuals. If the residuals are spread equally around the predictors, the model fulfills the assumption of homoskedasticity. In this case, it can be argued that there is an increase in the spread of the residuals which indicates heteroskedasticity.

The residuals vs leverage plot helps detect outliers in the model. It is important because outliers can have a major impact in the model. The residuals can be seen if they have a large Cook's distance, which is the dashed line. In this case, it does not seem like there is any observations with a large enough Cook's distance to be outside of the limits. There are some observations quite far apart from the rest but none outside of the limits from the dashed lines.

Model 2

```
par(mfrow = c(2, 2))
plot(model2)
```



From the residuals vs fitted plot it can be seen that the residuals are more equally spread around the line than in the first model. The line also seems to be more horizontal where the first model showed a more decreasing line.

From the Q-Q residuals plot it can be seen that the residuals fit the dotted line better than the first model although still not perfectly. This indicates that the residuals are approximately normally distributed and at least closer to a normal distribution than the first model.

The scale-location plot does not show the same signs of heteroskedasticity as in the first model. The residuals are closer together and are spread more equally around the horizontal line in this model.

The residuals vs leverage plot does not show any significant outliers as none of the observations have a large enough Cook's distance to be outside of the limits.

Because it fits the four models the best and has the highest adjusted R^2, it can be argued that model 2 would be preferred over model 1.

2.3. Examine whether the two models are misspecified using the RESET test.

The null hypothesis, H_0 is that the model is correctly specified:

$$H_0 : \delta_1 = 0 =, \delta_2 = 0$$

And the alternative hypothesis is:

$$H_1 : \delta_1 \neq 0, \delta_2 \neq 0$$

First we use the values found from the two models to find our $\hat{y}$ for model 1 and model 2. Setting up the new models to perform a RESET test:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + v$$

where $\hat{y}$ is the fitted value from the original model. $\hat{y}^2$ and $\hat{y}^3$ will capture any nonlinearities, if present in the model.

```
yhat1 <- -6.93228 + 0.99327*educ + 1.60816*salbegin + 1.83088*male - 1.72539*minority #model 1
yhat2 <- 0.849130 + 0.023578*log(salbegin) + 0.045474*male - -0.041856*minority #model 2
yhat1_sq <- yhat1^2; yhat1_cub <- yhat1^3
yhat2_sq <- yhat2^2; yhat2_cub <- yhat2^3
mod1_res <- lm(salary ~ educ + salbegin + male + minority + yhat1_sq + yhat1_cub)
mod2_res <- lm(log(salary) ~ educ + log(salbegin) + male + minority + yhat2_sq + yhat2_cub)
```

When the the RESET models has been made, we can use waldtest to find the F-statistics:

```
waldtest(mod1_res, terms=c("yhat1_sq","yhat1_cub"))
```

```
## Wald test
##
## Model 1: salary ~ educ + salbegin + male + minority + yhat1_sq + yhat1_cub
## Model 2: salary ~ educ + salbegin + male + minority
##   Res.Df Df      F  Pr(>F)
## 1    443
## 2    445 -2 2.5756 0.07725 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For model 2:

```
waldtest(mod2_res, terms=c("yhat2_sq","yhat2_cub"))
```

```
## Wald test
##
## Model 1: log(salary) ~ educ + log(salbegin) + male + minority + yhat2_sq +
##     yhat2_cub
## Model 2: log(salary) ~ educ + log(salbegin) + male + minority
##   Res.Df Df     F Pr(>F)
## 1    443
## 2    445 -2 0.331 0.7184
```

To confirm our results, we can use (resettest)

```
resettest(model); resettest(model2)
```

```
##
##  RESET test
##
## data:  model
## RESET = 2.5756, df1 = 2, df2 = 443, p-value = 0.07725
```

```
##
##  RESET test
##
## data:  model2
## RESET = 2.6338, df1 = 2, df2 = 443, p-value = 0.07293
```

Since both models have a p-value higher than 5%, respectively 7.7% and 7.3%, H_0 cannot be rejected, and therefore we cannot conclude that any of the models are misspecified. However, both p-values are still pretty low, and would not be accepted at e.g. a 10% significance level.

2.4. Explain why it could be relevant to include $educ^2$ as an explanatory variable in the two models. Estimate the two models again with $educ^2$ included (along with its corresponding coefficient β_5). Briefly comment on the output, and perform the RESET test again.

If it is still assumed, that the model is mis-specified, then it could be argued that including the squared could help correctly specifying the model. Excluding the variable could lead to a bias on educ in the original model. By excluding the variable $educ^2$, it would be correlated with the error term, why the model would not fulfill assumption MLR4. Furthermore, by including $educ^2$, the model will be able to capture any quadratic forms.

Model 1

```
model_ed <- lm(salary~educ+salbegin+male+minority+I(educ^2))
summary(model_ed)
```

```
##
## Call:
## lm(formula = salary ~ educ + salbegin + male + minority + I(educ^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.264  -4.042  -0.870   2.891  49.720
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 14.60237    6.78940   2.151              0.03203 *
## educ        -2.30474    1.01458  -2.272              0.02359 *
## salbegin     1.47994    0.07438  19.897 < 0.0000000000000002 ***
## male         1.78553    0.84791   2.106              0.03578 *
## minority    -1.61496    0.91115  -1.772              0.07701 .
## I(educ^2)    0.13205    0.04008   3.294              0.00107 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.79 on 444 degrees of freedom
## Multiple R-squared:  0.8011, Adjusted R-squared:  0.7989
## F-statistic: 357.7 on 5 and 444 DF,  p-value: < 0.00000000000000022
```

Model 2

```
model_ed2 <- lm(log(salary)~educ+log(salbegin)+male+minority+I(educ^2))
summary(model_ed2)
```

```
##
## Call:
## lm(formula = log(salary) ~ educ + log(salbegin) + male + minority +
##     I(educ^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44932 -0.11908 -0.00702  0.11246  0.87400
##
## Coefficients:
##                 Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    1.1754018  0.2073494   5.669         0.0000000259 ***
## educ          -0.0147950  0.0229937  -0.643                0.5203
## log(salbegin)  0.7825191  0.0433061  18.069 < 0.0000000000000002 ***
## male           0.0483029  0.0207975   2.323                0.0207 *
## minority      -0.0416353  0.0210129  -1.981                0.0482 *
## I(educ^2)      0.0015510  0.0009153   1.694                0.0909 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1782 on 444 degrees of freedom
## Multiple R-squared:  0.8064, Adjusted R-squared:  0.8042
## F-statistic: 369.8 on 5 and 444 DF,  p-value: < 0.00000000000000022
```

```
resettest(model_ed); resettest(model_ed2)
```

```
##
##  RESET test
##
## data:  model_ed
## RESET = 1.8771, df1 = 2, df2 = 442, p-value = 0.1543

##
##  RESET test
##
## data:  model_ed2
## RESET = 1.8884, df1 = 2, df2 = 442, p-value = 0.1525
```

Here it can be seen that the p-value increases, when $educ^2$ is included, and therefore we can still not reject H_0. The non-linear forms of the fitted values are not statistically significant.

We can set up comparisons of the two models, with and without $educ^2$

```
screenreg(list(Model1=model, Model1_ed2=model_ed))
```

```
##
## ====================================
```

```
##              Model1      Model1_ed2
## ----------------------------------
## (Intercept)   -6.93 ***   14.60 *
##               (1.86)      (6.79)
## educ           0.99 ***   -2.30 *
##               (0.17)      (1.01)
## salbegin       1.61 ***    1.48 ***
##               (0.06)      (0.07)
## male           1.83 *      1.79 *
##               (0.86)      (0.85)
## minority      -1.73       -1.61
##               (0.92)      (0.91)
## educ^2                     0.13 **
##                           (0.04)
## ----------------------------------
## R^2            0.80        0.80
## Adj. R^2       0.79        0.80
## Num. obs.    450         450
## ==================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

In the above lists, there is an upward bias on educ, when $educ^2$ is not included, whereas there is a negative effect when $educ^2$ is included, while education in the long run will affect your salary positive.

Comparing for log(salary):

```
screenreg(list(Model2=model2, Model2_ed2=model_ed2), digits=5)
```

```
##
## ==========================================
##                Model2         Model2_ed2
## ------------------------------------------
## (Intercept)      0.84913 ***    1.17540 ***
##                 (0.07709)      (0.20735)
## educ             0.02358 ***   -0.01480
##                 (0.00399)      (0.02299)
## log(salbegin)    0.82073 ***    0.78252 ***
##                 (0.03705)      (0.04331)
## male             0.04547 *      0.04830 *
##                 (0.02077)      (0.02080)
## minority        -0.04186 *     -0.04164 *
##                 (0.02106)      (0.02101)
## educ^2                          0.00155
##                                (0.00092)
## ------------------------------------------
## R^2              0.80510        0.80636
## Adj. R^2         0.80335        0.80418
## Num. obs.      450            450
## ==========================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

For these, the same pattern can be observed. Education will affect negatively for short education, while a long education will affect your salary positively. Both models has a convex shape, where education will

affect negatively on the short run, but there is an accelerating effect, as you keep studying. By performing RESET tests for the models including $educ^2$, we can find out whether the models are better specified.

```
resettest(model_ed);resettest(model_ed2)
```

```
##
##  RESET test
##
## data:  model_ed
## RESET = 1.8771, df1 = 2, df2 = 442, p-value = 0.1543
```

```
##
##  RESET test
##
## data:  model_ed2
## RESET = 1.8884, df1 = 2, df2 = 442, p-value = 0.1525
```

Since the p-values increase, the models are now better specified, and H_0 cannot be rejected.

2.5. Test the hypothesis $H_0 : \beta_1 = \beta_5 = 0$ in both models (from question 4).

To test the hypothesis $H_0 : \beta_1 = \beta_5 = 0$ in both models from question 4, we can use the F-test. This tests whether both the coefficient for educ and the coefficient for $educ^2$ are equal to zero. If the null hypothesis is rejected, it indicates that at least one of these coefficients are statistically significant.

The formula for the F-statistics is:

$$F = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} * \frac{n - k - 1}{q}$$

where R_{ur}^2 is the R-squared from our unrestricted model, R_r^2 is the R-squared from the restricted model, q is the difference in the degrees of freedom between the unrestricted and restricted model, n is the number of observations in the dataset and k is the number of independent variables in the unrestricted model.

The two unrestricted models are given in 2.4, so what we need to do is estimate our two restricted models, that looks like the following:

$$salary = \beta_0 + \beta_2 salbegin + \beta_3 male + \beta_4 minority + u$$

$$log(salary) = \beta_0 + + \beta_2 log(salbegin) + \beta_3 male + \beta_4 minority + u$$

```
model_r <- lm(salary ~ salbegin + male + minority, data = data2)
model_r2 <- lm(log(salary) ~ log(salbegin) + male + minority, data = data2)
```

We then obtain our R^2 from the restricted and unrestricted models:

```
r2_ur <- summary(model_ed)$r.squared
r2_r <- summary(model_r)$r.squared

r2_ur2 <- summary(model_ed2)$r.squared
r2_r2 <- summary(model_r2)$r.squared
```

We now have what we need to calculate our F-statistics:

```
F1 <- (r2_ur-r2_r)/(1-r2_ur) * (450-5-1)/2
F2 <- (r2_ur2-r2_r2)/(1-r2_ur2) * (450-5-1)/2
F1
```

```
## [1] 23.56319
```

```
F2
```

```
## [1] 18.9423
```

We then calculate our critical values of the F-statistics at 5% significance level:

```
qf(0.95, 2, 450-5-1)
```

```
## [1] 3.016036
```

Our decision rule says that if F > c then we reject our null hypothesis and instead accept our alternative hypothesis meaning that β_1 and β_5 for the education does have a statistical impact on yearly salary in both of our models.

It is also possible to do the test directly in R by using the following code:

```
myh0 <- c("educ=0", "I(educ^2)=0")
linearHypothesis(model_ed, myh0)
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ = 0
## I(educ^2) = 0
##
## Model 1: restricted model
## Model 2: salary ~ educ + salbegin + male + minority + I(educ^2)
##
##   Res.Df   RSS Df Sum of Sq      F         Pr(>F)
## 1    446 29801
## 2    444 26941  2    2859.5 23.563 0.000000000188 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(model_ed2, myh0)
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ = 0
## I(educ^2) = 0
##
## Model 1: restricted model
## Model 2: log(salary) ~ educ + log(salbegin) + male + minority + I(educ^2)
##
```

```
##   Res.Df    RSS Df Sum of Sq      F        Pr(>F)
## 1    446 15.306
## 2    444 14.102  2    1.2033 18.942 0.00000001275 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the code in R to perform the F-statistics gives us the same answer as before still confirming the rejecting of our null hypothesis.

2.6. Could there be issues with measurement errors in the two models? In what cases would it pose a problem?

It is possible that there could be some kind of measurement errors in the two models. Sometimes, it is difficult to collect data that truly reflects the correct value. If we think of the dependent variable in this model, which is income, then there can be different problems depending on how the data is collected. If it is collected from the tax register, then it might not reflect the true value of people's income, because some might have a larger income if they have done some kind of undeclared work that is not registered. If the data is obtained by asking people about their income then there might be a difference between what they tell and their actual income.

There can also be some problems with the independent variables in the model. Some people might lie about their education, or some can have problems remembering how many years of experience they have. For all variables that is included and especially data that reflects economic behavior it can be very difficult to obtain true and reliable data and avoid the above mentioned measurement errors.

We can look at the two different scenarios and what kind of effects it can have in the model, when there is a measurement error in the dependent variable or in the independent variable.

Measurement errors in dependent variable

We assume the following model has measurement errors in the dependent variable:

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

Where the measurement error is given by difference between the observed y^* and the true y:

$$e_0 = y - y^*$$

So we can write:

$$y^* = y - e_0$$

If we substitute y^* into the original model we get:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + (u + e_0)$$

So we can now see that measurement errors will affect the error term in the model, which will increase the variance of the model and have an effect on the coefficients variance and thereby the estimates of the population values. But is the estimator biased or unbiased in the case of measurement errors? We can say that the OLS estimators are unbiased and consistent if the measurement error in the dependent variable is statistically independent of the explanatory variables under the assumptions:

$$E(u) = E(e_0) = 0 \qquad Cov(u, e_0) = 0, \quad Cov(e_0, x) = 0$$

If e_0 does not have a zero mean, then we will get a biased estimate of the intercept, but the slope would remain the same. If e_0 and u are uncorrelated then the variance can be calculated as:

$$Var(u + e_0) = \sigma_u^2 + \sigma_e^2$$

Measurement errors in independent variable

If we look at a model where we assume measurement errors in one of the independent variables:

$$y = \beta_0 + \beta_1 x_1^* + u$$

The measurement error in x_1^* is given by $e_1 = x_1 - x_1^*$, where we can isolate the observed variable x_1^*, so $x_1^* = x_1 - e_1$. We can substitute this into the original model:

$$y = \beta_0 + \beta_1(x_1 - e_1) + u$$

$$y = \beta_0 + \beta_1 x_1 + u - \beta_1 e_1$$

, where $E(u) = E(e) = 0$

How measurement errors in independent variables affect OLS estimates depends upon which assumptions you make, and usually there are two extreme assumptions discussed.

Assumption 1 - uncorrelated with x_1

The first assumption is that the measurement error is uncorrelated with the true value x_1 and if this is true then the measurement error must be correlated with the observed value x_1^*:

$$Cov(x_1, e_1) = 0 \qquad Cov(x_1^*, e_1) \neq 0$$

In this case the regression $y = \beta_0 + \beta_1 x_1 + u - \beta_1 e_1$ will be unbiased because it was assumed that the measurement error was not correlated with x_1. The estimates of the model will therefore still be consistent and unbiased, but as in the case with measurement error in the dependent variable it will lead to an increase in the variance of the model, because the error term is changed:

$$Var(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_e^2$$

Assumption 2 - uncorrelated with x_1^*

The other assumption is when the measurement error is uncorrelated with the observed value x_1^* meaning that it must be correlated with the true value x_1:

$$Cov(x_1, e_1) \neq 0 \qquad Cov(x_1^*, e_1) = 0$$

In econometrics this particular case is known as classical errors-in-variables (CEV). In this case it would create biased estimators because the measurement error is correlated with the error term which means that x_1 is also assumed to be correlated with the error term of the model, making it biased and inconsistent.

It should be noted that a measurement error in one of the independent variable will cause bias and inconsistency in all the other included independent variables under assumption 2.

Exam 3 - Instrumental variables

In a multiple linear regression (MLR) there are 6 assumptions. Assumption 1-5 are called Gauss-Markov assumptions and assumption 6 is called the normality assumption. The first 4 assumptions exist to secure that the model is unbiased, while assumption 5 checks for heteroskedasticity and assumption 6 checks for normality in the model. The assumptions are:

MLR1: Linear in Parameters

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters of interest and u is an unobserved random error or disturbance term.

MLR2: Random Sampling

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

MLR3: No Perfect Collinearity

In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables.

MLR4: Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0$$

MLR5: Homoskedasticity

The error u has the same variance given any value of the explanatory variables. In other words,

$$\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$$

MLR6: Normality

The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0, \sigma^2)$.

Consider the following model: $log(earning) = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 male + \beta_4 ethblack + \beta_5 ethhisp + u$

where earnings are hourly wages in US dollars, educ is education measured in years of schooling, exp is work experience measured in years, male is a gender dummy, ethblack and ethhisp are race dummies for African Americans and Hispanics, respectively. Additionally, we have three instruments: mother's education measured in years (meduc), father's education measured in years (feduc), and number of siblings (sibling s).

3.1. Estimate the model using OLS and comment on the results.

Here we use the lm() function to estimate the model.

```
model <- lm(log_earning~educ+exp+male+black+hisp, , data = data3)
summary(model)
```

```
##
## Call:
## lm(formula = log_earning ~ educ + exp + male + black + hisp,
##     data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07585 -0.28006 -0.00145  0.30775  1.98441
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  0.396226   0.173508   2.284              0.02280 *
## educ         0.124220   0.009451  13.143 < 0.0000000000000002 ***
## exp          0.033882   0.005046   6.715      0.0000000000499 ***
## male         0.293449   0.045803   6.407      0.0000000003363 ***
## black       -0.195670   0.071255  -2.746              0.00624 **
## hisp        -0.097406   0.100342  -0.971              0.33213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5103 on 514 degrees of freedom
## Multiple R-squared:  0.3539, Adjusted R-squared:  0.3476
## F-statistic: 56.32 on 5 and 514 DF,  p-value: < 0.00000000000000022
```

From the model, it can be seen that the education variable has an estimate of 0.124220 which means that one extra year of education will raise the hourly wage by approximately 12.42%. The experience variable has an estimate of 0.033882 which means that one more year of experience will raise the hourly wage by approximately 3.39%. The male variable has an estimate of 0.293449 which means that if you are a male your hourly wage will be approximately 29.34% higher than if you are a woman. The black variable has an estimate of -0.19567 which means that African Americans approximately will have 19.57% lower hourly wages than non-African Americans. The Hispanics variable has an estimate of -0.097406 which means that Hispanic approximately will have 9.74% lower hourly wages than non-Hispanics.

The intercept is 0.396226 which is the value of the dependent variable, in this case hourly wage in US dollars, when all other variables have a value of 0.

It can also be seen that all the variables except the hispanic variable are statistically significant at a 5% significance level due to p-values < 0.05. They are also significant at a 1% significance level.

3.2. Why might we be concerned that education is endogenous?

Given the significant positive relationship between education and earnings, it is important to consider if endogeneity is biasing the results. The potential endogeneity of education needs to be tested through instrumental variables (IV) regression. There may be factors that affect both education and earnings that are not included in the model. For example family background could influence both education and earning potential. If these factors are not accounted for, the estimated coefficient on education will capture not only the effect of education but also the effect of these omitted variables, leading to bias.

3.3. Are sibling, meduc, and feduc useful as instruments?

To find out whether these variables are useful instruments, they have to fulfill two conditions; they need to be correlated with educ and they need to be independent of the error term, u:

$$Cov(x, z) \neq 0$$

$$Cov(z, u) = 0$$

In this case, x is the variable, z is the instruments and u is the error term in the model. If the instruments are correlated with the error term, these are also affected by an endogeneity issue. The instruments need to be uncorrelated with the omitted variable, that creates an endogeneity issue for educ. The first condition, $Cov(x, z) \neq 0$, can be tested. A model for the variable educ as the dependent variable can be set up. When multiple IVs are used, 2SLS is used:

$$educ = \pi_0 + \pi_1 exp + \pi_2 male + \pi_3 black + \pi_4 hisp + \pi_5 sib + \pi_6 meduc + \pi_7 feduc + v$$

Setting up the hypothesis test:

$$H_0 : \pi_5 = 0, \pi_6 = 0 \; or \; \pi_7 = 0$$

$$H_1 : \pi_5 \neq 0, \pi_6 \neq 0 \; or \; \pi_7 \neq 0$$

By using F-tests, we can test the hypothesis. F-stat is calculated using the following formula:

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} * \frac{n - k - 1}{q}$$

```r
sib <- data3$siblings; meduc <- data3$meduc; feduc <- data3$feduc
ivreg <- lm(educ~exp+male+black+hisp+sib+meduc+feduc)
summary(ivreg)
```

```
##
## Call:
## lm(formula = educ ~ exp + male + black + hisp + sib + meduc +
##     feduc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8750  -1.3575  -0.2666   1.4341   8.9572
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 11.240160   0.653682  17.195 < 0.0000000000000002 ***
## exp         -0.116651   0.020372  -5.726         0.0000000175 ***
## male         0.003102   0.190843   0.016               0.9870
## black       -0.631488   0.299167  -2.111               0.0353 *
## hisp         0.658727   0.438636   1.502               0.1338
## sib         -0.102192   0.046651  -2.191               0.0289 *
## meduc        0.261472   0.047138   5.547         0.0000000467 ***
## feduc        0.147712   0.035030   4.217         0.0000293049 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.118 on 512 degrees of freedom
## Multiple R-squared:  0.2778, Adjusted R-squared:  0.2679
## F-statistic: 28.13 on 7 and 512 DF,  p-value: < 0.00000000000000022
```

It can be seen, that all three factors are significant at a 5% significance level.

```r
ivtest <- c("sib=0","meduc=0","feduc=0")
linearHypothesis(ivreg,ivtest)
```

```
## Linear hypothesis test
##
## Hypothesis:
## sib = 0
## meduc = 0
## feduc = 0
##
## Model 1: restricted model
## Model 2: educ ~ exp + male + black + hisp + sib + meduc + feduc
##
##   Res.Df    RSS Df Sum of Sq      F               Pr(>F)
## 1    515 2914.6
## 2    512 2297.8  3    616.75 45.809 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value < 0.05 H_0 is rejected, hence the three variables are statistically significant. Thereby it can be assumed, that the three variables are correlated with education, and can be used as instruments for education.

3.4. Test whether education is endogenous.

According to Hausman, the method to test whether a variable is endogenous, is to examine if the results of OLS and 2SLS are different, and if this difference is statistically significant. If it is stastistically significant, it can be argued that there is endogeneity. To test this, we find the reduced form residuals from the linear regression including the three Instrument Variables. The reduced form residuals is denoted as $\hat{v}$. $\hat{v}$ is then included in the original OLS model, and if the residuals are significant, it means that there is an endogeneity issue:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \delta v + error$$

Thereby we get the nullhypothesis:

$$H_0 : \delta = 0$$

```
res <- ivreg$residuals
res_mod <- lm(log_earning~educ+exp+male+black+hisp+res)
coeftest(res_mod)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value           Pr(>|t|)
## (Intercept) -0.064700   0.339003 -0.1909            0.84871
## educ         0.153036   0.020516  7.4593 0.0000000000003742 ***
## exp          0.037628   0.005567  6.7591 0.0000000000378421 ***
## male         0.290479   0.045775  6.3458 0.0000000004868537 ***
## black       -0.157544   0.075122 -2.0972            0.03647 *
## hisp        -0.069476   0.101739 -0.6829            0.49499
## res         -0.036550   0.023106 -1.5818            0.11430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since residuals are not significant, we fail to reject H_0, and thereby we cannot conclude that there is an endogeneity issue. This does not mean that educ does not have an endogeneity issue, but for these IV's there is not.

3.5. Estimate the model using 2SLS employing the three described instruments. Compare with the results in question 1.

The p-value of 0.11 is not sufficient to reject that there is an endogeneity issue, why it is deemed necessary to further examine if there is an issue. To accomodate this issue, we use the 2SLS method. A regression on this form is set up:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 exp + \beta_3 male + \beta_4 black + \beta_5 hisp + u$$

where y_2=educ, the variable that is suspected to have endogeneity. Now, a regression for educ is made. The three IV's for educ is added in the regression:

$$y_2 = \pi_0 + \pi_1 exp + \pi_2 male + \pi_3 black + \pi_4 hisp + \pi_5 sib + \pi_6 meduc + \pi_7 feduc + v$$

In this regression, the following assumptions is expected to be fulfilled:

- $E(v) = 0$. The mean of the error term is 0, which indicates unbiasedness.
- $Cov(x, v) = 0$. This applies for all variables in the original model.

- $Cov(z, v) = 0$. The IV's should be independent of the error term.

```
stage1 <- lm(educ~exp+male+black+hisp+sib+meduc+feduc)
educ_fitted <- fitted(stage1)
```

As seen in 3.3 the IV's are significant variables for education:

```
linearHypothesis(stage1,c("meduc=0", "feduc=0", "sib=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## meduc = 0
## feduc = 0
## sib = 0
##
## Model 1: restricted model
## Model 2: educ ~ exp + male + black + hisp + sib + meduc + feduc
##
##   Res.Df    RSS Df Sum of Sq      F                Pr(>F)
## 1    515 2914.6
## 2    512 2297.8  3    616.75 45.809 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value < 0.05, H_0 is rejected, and the IV's are significant. The fitted values of educ is used as IV. A new expression for y_2 is found: $y_2 = \hat{y_2} + v$ This new expression can be added to the original regression:

$$y_1 = \beta_0 + \beta_1(\hat{y_2} + v) + \beta_2 exp + \beta_3 male + \beta_4 black + \beta_5 hisp + u$$

```
stage2 <- lm(log_earning~educ_fitted+exp+male+black+hisp)
sls <- ivreg(log_earning~educ+exp+male+black+hisp|exp+male+black+hisp+sib+feduc+meduc)
```

Setting the new 2SLS and the OLS up:

```
screenreg(list(OLS=model,TwoSLS=sls), digits=4)
```

```
##
## ======================================
##              OLS          TwoSLS
## --------------------------------------
## (Intercept)    0.3962 *     -0.0647
##               (0.1735)      (0.3426)
## educ           0.1242 ***    0.1530 ***
##               (0.0095)      (0.0207)
## exp            0.0339 ***    0.0376 ***
##               (0.0050)      (0.0056)
## male           0.2934 ***    0.2905 ***
##               (0.0458)      (0.0463)
## black         -0.1957 **    -0.1575 *
##               (0.0713)      (0.0759)
## hisp          -0.0974       -0.0695
##               (0.1003)      (0.1028)
## --------------------------------------
## R^2            0.3539        0.3422
## Adj. R^2       0.3476        0.3358
## Num. obs.    520           520
## ======================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

It can be seen, that when using the 2SLS, the estimate for educ increases a bit, and Afro-American as well as Hispanic falls. Male and experience does not really change.

3.6. Perform the overidentification test. What do you conclude?

For a single endogenous variable, we need at least one instrument, but if we want to use more than one instrument for one endogenous variable, then we risk overidentifying. We used 3 instruments to estimate educ in 3.3. Because we have more instruments than needed we can perform an overidentification test to see whether some of them are uncorrelated with the structural error. The purpose of the test is to test whether the difference on the 3 coefficients are statistically significant. Is this the case, then it can be concluded that one of the instruments (or all) is showing endogeneity issues.

Just because they return statstically similar estimates, that does not mean that they necessarily are exogenous and does not have endogeneity issues and is correlated with the error term. If it can be concluded that they are statistically significant different, then some of the instruments needs to be removed.

First, the 2SLS residuals, $\hat{u}$, from the earlier estimated 2SLS model are obtained.

```
resid2sls <- resid(sls)
```

Then the residuals are regressed on all exogenous variables including the IV's, where R_1^2 is obtained.

```
res.aux <- lm(resid2sls ~ exp + male + black + hisp + sib + meduc + feduc)
```

The null hypothesis is that all IV's are uncorrelated with u:

$$H_0 : \text{Corr}(Z, u) = 0, \quad \text{where} \quad Z = (z_1, z_2, z_3).$$

We use $nR_1^2 \sim \chi_q^2$, where q is the number of instrumental variables from outside the model minus the total number of endogenous explanatory variables. Hypothesis will be tested:

```
r2 <- summary(res.aux)$r.squared
n <- nobs(res.aux)
teststat <- n*r2
pval <- 1-pchisq(teststat,df=2)
pval
```

```
## [1] 0.01455821
```

From the calculated p-value, H_0 is rejected at a 5% significance level due to the p-value < 0.05. This means that at least some of the IV's are not exogenous, which means that the model does not pass the test and is assumed to have overidentification issues.

3.7. Perform the entire analysis again using only meduc and feduc as instruments. Does this change your conclusions?

We perform the entire analysis again, starting from 3.3, because 3.1 and 3.2 are not about the choice of instrumental variables. The tests and methods will not be elaborated further, so for an explanation of what we are doing look at the earlier assignments.

We start by testing whether meduc and feduc are useful as instruments using the same method as in 3.3:

```
ivreg7 <- lm(educ ~ exp + male + black + hisp + meduc + feduc)
linearHypothesis(ivreg7, c("meduc=0","feduc=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## meduc = 0
## feduc = 0
##
## Model 1: restricted model
## Model 2: educ ~ exp + male + black + hisp + meduc + feduc
##
##   Res.Df    RSS Df Sum of Sq      F                Pr(>F)
## 1    515 2914.6
## 2    513 2319.3  2    595.22 65.826 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value < 0.05 H_0 is rejected, the two variables are statistically significant, why it can be assumed, that meduc and feduc are correlated with education, and can be used as instruments for education.

Now we want to test if there is endogeneity issues in the new model with the two instrumental variables. The same method as in 3.4 will be used:

```
res7 <- ivreg7$residuals
res_mod7 <- lm(log_earning ~ educ + exp + male + black + hisp + res7)
coeftest(res_mod7)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value           Pr(>|t|)
## (Intercept) -0.1657973  0.3443473 -0.4815            0.63038
## educ         0.1593561  0.0208626  7.6384 0.0000000000001086 ***
## exp          0.0384495  0.0055843  6.8853 0.0000000000169123 ***
## male         0.2898274  0.0457297  6.3378 0.0000000005107438 ***
## black       -0.1491819  0.0752218 -1.9832            0.04787 *
## hisp        -0.0633502  0.1017049 -0.6229            0.53364
## res7        -0.0441531  0.0233868 -1.8879            0.05960 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We fail to reject the null hypothesis at a 5% significance level, and thereby we cannot conclude that there is an endogeneity issue. If we used a 10% significance level then we could have rejected the null hypothesis and instead accepted the alternative hypothesis that there is an endogeneity issue, indicating potential endogeneity issues.

We perform the 2SLS method again as in 3.5, but this time with only meduc and feduc as instrumental variables, and we do it directly with the code in R:

```
sls2 <- ivreg(log_earning ~ educ + exp + male + black + hisp | exp + male + black + hisp + feduc + medu
```

To see whether it changes our conclusions we compare it with the OLS from 3.1 and the 2SLS from 3.5:

```
screenreg(list(OLS=model, SLS3_5 = sls, SLS3_7=sls2), digits = 4)
```

```
##
## ===================================================
##              OLS           SLS3_5        SLS3_7
## ---------------------------------------------------
## (Intercept)    0.3962 *     -0.0647       -0.1658
##               (0.1735)      (0.3426)      (0.3498)
## educ           0.1242 ***    0.1530 ***    0.1594 ***
##               (0.0095)      (0.0207)      (0.0212)
## exp            0.0339 ***    0.0376 ***    0.0384 ***
##               (0.0050)      (0.0056)      (0.0057)
## male           0.2934 ***    0.2905 ***    0.2898 ***
##               (0.0458)      (0.0463)      (0.0465)
## black         -0.1957 **    -0.1575 *     -0.1492
##               (0.0713)      (0.0759)      (0.0764)
## hisp          -0.0974       -0.0695       -0.0634
##               (0.1003)      (0.1028)      (0.1033)
## ---------------------------------------------------
## R^2            0.3539        0.3422        0.3366
## Adj. R^2       0.3476        0.3358        0.3301
## Num. obs.    520           520           520
## ===================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

In the above we can see the original model and the different results we get with using 3 IV's and 2 IV's to estimate the fitted values of education. We performed overidentification test in 3.6 when using 3 IV's,

where we found that the model was indeed overidentificated, so one of the IV's was not exogenous. The same overidentification test is now performed on the new model with two IV's:

```
resid2sls2 <- resid(sls2)
res.aux2 <- lm(resid2sls2 ~ exp + male + black + hisp + meduc + feduc)
r2_2 <- summary(res.aux2)$r.squared
n2 <- nobs(res.aux2)
teststat2 <- n2*r2_2
pval2 <- 1-pchisq(teststat2,df=1) #now only 1 degree of freedom, because only 2 IV's and 1 endogenous v
pval2
```

```
## [1] 0.01593513
```

The new model with 2 IV's also rejects the null hypothesis. This means that at least one of the IV's are not exogenous, which means that the model does not pass the test and is assumed to have overidentification.

Exam 4 - Models for binary variables

In a multiple linear regression (MLR) there are 6 assumptions. Assumption 1-5 are called Gauss-Markov assumptions and assumption 6 is called the normality assumption. The first 4 assumptions exist to secure that the model is unbiased, while assumption 5 checks for heteroskedasticity and assumption 6 checks for normality in the model. The assumptions are:

MLR1: Linear in Parameters

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters of interest and u is an unobserved random error or disturbance term.

MLR2: Random Sampling

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

MLR3: No Perfect Collinearity

In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables.

MLR4: Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0$$

MLR5: Homoskedasticity

The error u has the same variance given any value of the explanatory variables. In other words,

$$\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$$

MLR6: Normality

The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0, \sigma^2)$.

In this assignment, we investigate the factors that influence whether women in Switzerland participate in the labor force.

The dependent variable is participation, a binary variable that measures whether the person is part of the labor force. Additionally, we have seven explanatory variables: income that is not work-related measured in 1000 CHF (income), age (age), age squared (agesq), education measured in years (educ), number of children under 7 years old (youngkids), number of children over 7 years old (oldkids), and a dummy variable indicating whether the person is a foreigner (foreign).

The dataset data4 contains these variables measured for 872 Swiss women.

4.1. Set up a linear regression model for participation where you use the described explanatory variables.

(a) Estimate the model using OLS and comment on the results.

Here we use the lm() function to estimate the model.

```
model <- lm(part~inc+age+age_sq+educ+you+old+fore, , data = data)
summary(model)
```

```
##
## Call:
## lm(formula = part ~ inc + age + age_sq + educ + you + old + fore,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84324 -0.39866 -0.08992  0.42048  1.01049
##
## Coefficients:
##               Estimate Std. Error t value           Pr(>|t|)
## (Intercept) -0.3685651  0.2529630  -1.457            0.14548
## inc         -0.0035163  0.0007098  -4.954 0.0000008754069771 ***
## age          0.0633852  0.0128603   4.929 0.0000009919903481 ***
## age_sq      -0.0009029  0.0001566  -5.767 0.0000000112290065 ***
## educ         0.0067725  0.0059615   1.136            0.25626
## you         -0.2390033  0.0313780  -7.617 0.0000000000000682 ***
## old         -0.0474930  0.0171593  -2.768            0.00576 **
## fore         0.2572106  0.0401252   6.410 0.0000000002387949 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4506 on 864 degrees of freedom
## Multiple R-squared:  0.1901, Adjusted R-squared:  0.1836
## F-statistic: 28.98 on 7 and 864 DF,  p-value: < 0.00000000000000022
```

From the model, it can be seen that the income variable has an estimate of -0.0035 which means that one unit (1000 CHF) more income will give approximately 0.35% lower probability of participation.

The age variable has an estimate of 0.0634 while the age squared variable has an estimate of -0.0009 which means that the probability of participation will be higher up to a certain age where the probability will start to drop. This makes sense due to people retiring at a certain age.

The education variable has an estimate of 0.0068 which means that one extra year of education will give approximately a 0.68% higher probability of participation.

The young kids variable for number of children under 7 years old has an estimate of -0.239 which means that if you have children under 7 years old, you will have approximately 23.9% lower probability of participation than people without children under 7 years old.

The old kids variable for number of children over 7 years old has an estimate of -0.0475 which means that if you have children over 7 years old, you will have approximately 4.75% lower probability of participation than people without children over 7 years old.

The foreign variable has an estimate of 0.2572 which means that you have approximately 25.72% higher probability of participation if you are foreign than if you are not foreign.

The intercept is -0.0035 which is the value of the dependent variable, in this case participation in the labor force, when all other variables have a value of 0.

It can also be seen that all the variables except education are statistically significant at a 5% significance level due to p-values < 0.05. All variables except education are also significant at a 1% significance level.

(b) Test whether the partial effect of education is different from zero.

Using a t-test, we can test whether the partial effect of education is different from zero. We set up our null hypothesis:

$$H_0 : \beta_4 = 0$$

And the alternative hypothesis:

$$H_1 : \beta_4 \neq 0$$

If the t-value is within the 95% accept-interval, we fail to reject our null hypothesis. Otherwise, we reject our null hypothesis, and instead accept our alternative hypothesis. We perform a two-sided test, which means there is 2.5% outside of the acceptance interval on each side.

The t-score can be calculated as follows:

$$t = \frac{\beta_j}{se(\beta_j)}$$

The coefficient, β is given from the summary of the model. The same goes for the standard error.

```
beta <-  0.0067725; se <- 0.0059615
tscore <- beta/se
tscore
```

```
## [1] 1.13604
```

To calculate the accept interval, we use the qt() function and our degrees of freedom

```
df <- length(data$educ)-1
acc <- qt(1-0.025, df)
interval <- c(-acc,acc)
names(interval) <- c("Lower limit", "Upper limit")
interval
```

```
## Lower limit Upper limit
##   -1.962691    1.962691
```

Since the calculated t-score (1.136) is within the accept interval, we fail to reject our null hypothesis at a 5% significance level.

(c) Test whether the partial effect of age is different from zero.

We can test whether the partial effect of age is different from zero. We set up our null hypothesis:

$$H_0 : \beta_2 = 0, \beta_3 = 0$$

And the alternative hypothesis:

$$H_1 : \beta_2 \neq 0, \beta_3 \neq 0$$

Since this is a joint hypothesis, an F-test is needed. To do this, we use the Wald-test. The Wald-test is a kind of F-test, that takes heteroskedasticity into account. The F-statistic is calculated using the following formula:

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} * \frac{n - k - 1}{q}$$

```
waldtest(model, vcov=vcovHC(model,type="HC0"), terms=(2:3))
```

```
## Wald test
##
## Model 1: part ~ inc + age + age_sq + educ + you + old + fore
## Model 2: part ~ inc + educ + you + old + fore
##   Res.Df Df      F                Pr(>F)
## 1    864
## 2    866 -2 37.745 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the Wald test we get a p-value < 0.05, and we reject our null hypothesis, and instead accept our alternative hypothesis, H_1, that the partial effect of age is not equal to 0.

4.2. Set up both a logit and a probit model for participation where you use the described explanatory variables.

The purpose of the probit and logit models are to model the probability of "success" so the models should lie in the unit interval [0,1]. To make this happen the model has a general function G($\cdot$) that takes a value between 0 and 1:

$$P(y = 1|x) = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

The function can be written as a logit and probit model.

Logit: $G(z) = \tau(z) = \frac{exp(z)}{[1+exp(z)]}$

Probit: $G(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} exp(-z^2/2)du$

From these models, the partial effect of the independent variables can not be seen directly. The partial effect of a change in a variable can be found by taking the partial derivative of a function g:

$$\frac{\partial P(y = 1|x)}{\partial x_j} = g(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \cdot \beta_j$$

Therefore, the estimated partial effects of a roughly continous variable, x_j are:

$$\Delta \hat{P}(y = 1 \mid x) \approx \left[g(\hat{\beta}_0 + x\hat{\beta})\hat{\beta}_j \right] \Delta x_j$$

(a) Estimate the models.

The logit model and the probit model are estimated:

```
model_logit <- glm(part~inc+age+age_sq+educ+you+old+fore, family = binomial(link = "logit"))
model_probit <- glm(part~inc+age+age_sq+educ+you+old+fore, family = binomial(link = "probit"))
screenreg(list( Logit = model_logit, Probit = model_probit), digits = 4)
```

```
##
## ============================================
##                  Logit          Probit
## --------------------------------------------
## (Intercept)       -4.3864 ***    -2.6698 ***
##                   (1.3037)       (0.7756)
## inc               -0.0231 ***    -0.0138 ***
##                   (0.0048)       (0.0028)
## age                0.3295 ***     0.2000 ***
##                   (0.0679)       (0.0402)
## age_sq            -0.0047 ***    -0.0028 ***
##                   (0.0008)       (0.0005)
## educ               0.0386         0.0231
##                   (0.0302)       (0.0181)
## you               -1.1777 ***    -0.7103 ***
##                   (0.1718)       (0.1005)
## old               -0.2354 **     -0.1439 **
##                   (0.0846)       (0.0510)
## fore               1.1908 ***     0.7286 ***
##                   (0.2042)       (0.1214)
## --------------------------------------------
## AIC             1032.1533      1031.6526
## BIC             1070.3196      1069.8189
## Log Likelihood  -508.0766      -507.8263
## Deviance        1016.1533      1015.6526
## Num. obs.        872            872
## ============================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

As explained earlier, the estimates can not be interpreted directly without calculating the partial effect. From this model, it can be seen whether the variables are statisically significant and whether they have a positive or negative effect on the dependent variable.

It can be seen that the income variable has a negative effect in both models while it is also significant at a 1% significance level.

The age variable has a positive effect in both models while the age squared variable has a negative effect in both models. This could indicate, like in 4.1a that the effect becomes negative at a certain age. They are both significant at a 1% significance level.

The education variable is not statistically significant which means that we can not conclude on this variable.

The young kids variable has a negative effect in both models while it is also significant at a 1% significance level.

The old kids variable has a negative effect in both models while it is also significant at a 1% significance level.

The foreign variable has a positive effect in both models while it is also significant at a 1% significance level.

(b) Test whether the partial effect of education is different from zero.

To test whether the partial effect of educ is different from zero in the models, we set up a null hypothesis and an alternative hypothesis:

$$H_0 : \beta_4 = 0$$

$$H_1 : \beta_4 \neq 0$$

These are tested using a t-test, as in 4.1a.

```
screenreg(list(Logit=coeftest(model_logit), Probit=coeftest(model_probit)),digits=4)
```

```
##
## ====================================
##              Logit        Probit
## ------------------------------------
## (Intercept)  -4.3864 ***  -2.6698 ***
##              (1.3037)     (0.7756)
## inc          -0.0231 ***  -0.0138 ***
##              (0.0048)     (0.0028)
## age           0.3295 ***   0.2000 ***
##              (0.0679)     (0.0402)
## age_sq       -0.0047 ***  -0.0028 ***
##              (0.0008)     (0.0005)
## educ          0.0386       0.0231
##              (0.0302)     (0.0181)
## you          -1.1777 ***  -0.7103 ***
##              (0.1718)     (0.1005)
## old          -0.2354 **   -0.1439 **
##              (0.0846)     (0.0510)
## fore          1.1908 ***   0.7286 ***
##              (0.2042)     (0.1214)
## ====================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Since neither educ is significance for the logit or the probit, we fail to reject H_0. The fact that we fail to reject H_0 does not mean, that we can conclude, that the effect of education is different from zero. Hence, it is not sure, that educ has an effect on participation rate.

(c) Test whether the partial effect of age is different from zero using a likelihood-ratio test.

To perform a hypothesis test on multiple parameters (age and age^2), we can use the likelihood-ratio test.

$$LR = 2(L_{ur} - L_r)$$

The LR-test is similiar to the F-test, but the multiplication of 2 is needed so that the LR approximates a chi-square distribution

We set up our null hypothesis and alternative hypothesis, that is:

$$H_0 : \beta_2 = 0, \beta_3 = 0$$

$$H_1 : \beta_2 \neq 0, \beta_3 \neq 0$$

We already have our logit and probit model, so we only need to setup the logitR and probitR:

```
logitR <- glm(part~inc+educ+you+old+fore, family = binomial(link = "logit"))
probitR <- glm(part~inc+educ+you+old+fore, family = binomial(link = "probit"))
```

First we calculate the LR-score for logit:

```
lrlogit <- 2*(logLik(model_logit)-logLik(logitR))
pvallog <- pchisq(lrlogit,df=2, lower.tail=F)
pvallog
```

```
## 'log Lik.' 1 (df=8)
```

For logit H_0 is rejected since p-value < 0.05, and we instead accept our alternative hypothesis, that the partial effect of age is different from 0.

The same can be done for probit:

```
lrprobit <- 2*(logLik(model_probit)-logLik(probitR))
pvalprob <- pchisq(lrprobit, df=2, lower.tail=F)
pvalprob
```

```
## 'log Lik.' 1 (df=8)
```

The same result can be observed here; p-value < 0.05. Hence, we once again reject H_0 and instead accept our alternative hypothesis: age is different from zero.

4.3. We want to compare the partial effect of income across the models. Calculate the average partial effect (APE) and comment on the results.

To compare the partial effect of income across the models we use the APE (average partial effect) method on both the probit and logit model and then compare it with the LPM. We do this because it makes little sense to compare the direct coefficient estimates from a logit or probit model (or LPM). APE method is the closest to a comparable estimate between models that we find and it also can be compared with the LPM.

The APE calculates the partial effect for all observations and then takes the average of the effects:

$$\hat{APE} = \hat{\beta}_j \left[n^{-1} \sum_{i=1}^{n} g(x_i \hat{\beta}) \right]$$

This is calculated directly in R:

```
ape_model_logit <- logitmfx(model_logit, data = data, atmean = F)
ape_model_probit <- logitmfx(model_probit, data = data, atmean = F)
screenreg(list(APE_Logit=ape_model_logit, APE_Probit=ape_model_probit, LMP=model),digits=4)
```

```
##
## ============================================================
##                  APE_Logit      APE_Probit     LMP
## ------------------------------------------------------------
## inc                -0.0046 ***    -0.0046 ***   -0.0035 ***
##                    (0.0010)       (0.0010)      (0.0007)
## age                 0.0657 ***     0.0657 ***    0.0634 ***
##                    (0.0135)       (0.0135)      (0.0129)
## age_sq             -0.0009 ***    -0.0009 ***   -0.0009 ***
##                    (0.0002)       (0.0002)      (0.0002)
## educ                0.0077         0.0077        0.0068
##                    (0.0060)       (0.0060)      (0.0060)
## you                -0.2350 ***    -0.2350 ***   -0.2390 ***
##                    (0.0342)       (0.0342)      (0.0314)
## old                -0.0470 **     -0.0470 **    -0.0475 **
##                    (0.0169)       (0.0169)      (0.0172)
## fore                0.2889 ***     0.2889 ***    0.2572 ***
##                    (0.0462)       (0.0462)      (0.0401)
## (Intercept)                                     -0.3686
##                                                 (0.2530)
## ------------------------------------------------------------
## Num. obs.           872            872           872
## Log Likelihood   -508.0766      -508.0766
## Deviance         1016.1533      1016.1533
## AIC              1032.1533      1032.1533
## BIC              1070.3196      1070.3196
## R^2                                              0.1901
## Adj. R^2                                         0.1836
## ============================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

When looking at the estimate for income in the logit and probit model it can be seen that they are the same and that there is a larger average effect of change in income on the probability of participation than in the LPM. In the logit and probit model the effect of a one unit increase in income is a -0.46% change in the probability of participation, where it in the LPM is -0.35%. The consistency in the results across the models reinforces the robustness of that an increase in income is associated with a lower probability of participation.

4.4. We want to compare the partial effect of the foreign variable across the models. Calculate the Average Partial Effect (APE) and comment on the results.

APE is used to to examine a dummy variable, foreign. We calculate APE using the method for discrete variables.

$$\hat{APE} = n^{-1} \sum_{i=1}^{n} [G[\beta_0 + \beta_1 income + \beta_2 age + \beta_3 agesq + \beta_4 educ + \beta_5 youngkids + \beta_6 oldkids + \beta_7]$$
$$- G[\beta_0 + \beta_1 income + \beta_2 age + \beta_3 agesq + \beta_4 educ + \beta_5 youngkids + \beta_6 oldkids]]$$

So the APE when foreign (β_7) is included minus APE when foreign is excluded. For each observation (i), the estimated effect of changing foreign from 0 to 1 is calculated.

We start by setting up a matrix for all variables and define if a person is foreign(1) or not(0):

```
cdata <- cbind(1, as.matrix(data[,c("income","age","agesq","educ","youngkids","oldkids","foreign")]))
cdata1 <- cdata
cdata1[,8] <- 1
cdata2<-cdata
cdata2[,8] <- 0
```

Afterwards, the coefficients for both models is found

```
pcoef <- model_probit$coefficients
lcoef <- model_logit$coefficients
```

Now, APE can be calculated for probit and logit. Starting with probit:

```
probitAPE <- mean(pnorm(cdata1%*%pcoef)-pnorm(cdata2%*%pcoef))
probitAPE
```

```
## [1] 0.2493686
```

And for logit:

```
logitAPE <- mean(pnorm(cdata1%*%lcoef)-pnorm(cdata2%*%lcoef))
logitAPE
```

```
## [1] 0.3362055
```

It can be seen, that the probit APE is approximately 0.25, which is almost the same as for the LPM (0.26). However, for logit, it can be seen, that there is a quite big difference, as logit gives a result of 0.336, which indicates, that the partial effects for probit is almost the same as for LPM, where as for logit there is a difference. Which one is more correct, is hard to interpret.

4.5. Why is the Average Partial Effect (APE) preferred over the Partial Effect at the Average (PEA)?

The partial effect at the average (PEA) can be seen as the partial effect at the "average" of a variable. It is written like this:

$$P\hat{E}A_j = g(\bar{x}\hat{\beta})\hat{\beta}_j$$

A challenge can arise when using the PEA for a variable that can only take a value of 0 and 1. An example could be the foreign variable used earlier in this assignment. When a variable has a value of either 0 or 1 it would not make sense to calculate the partial effect of the average person as the PEA does.

On the other hand, the average partial effect (APE) calculates the partial effect for all observations and then takes the average of the effect. As shown earlier in this assignment, it is written like this:

$$A\hat{P}E = \hat{\beta}_j[n^{-1}\sum_{i=0}^{n} g(x_i\hat{\beta})]$$

Therefore, the reason that APE is preferred over PEA is that the APE shows the partial effect where as the PEA shows the partial effect of the average person.

4.6. Compare the models' predictive abilities by calculating the percent correctly predicted for each model.

Using a goodness of fit estimate called percently correctly predicted (PCP), which shows how big a share of the data that the models predicts correctly. However, PCP is not necessarily always a good indicator of the model. E.g. if 300 answers yes and 60 answers no, but the model predicts that everyone answers yes, it will have a high PCP, but is not necessarily a good predictor. To examine when the model predicts correctly, we set up a probability limit for fitted values of the models:

$$\tilde{y} = 1, if\ \tilde{y}\ \geq\ c$$

$$\tilde{y} = 0, if\ \tilde{y}\ \leq\ c$$

The probability limit is set to 0.5. Hence, fitted values of the three models below 0.5 will predict that the person is not participating in the labor force (y=0), while fitted values over 0.5 will predict, that the person participates in the labor force (y=1).

PCP is calculated in R in the following way:

```
y <- data$participation
lmp_pcp <- 100 * mean((model$fitted.values > 0.5) == y)
logit_pcp <- 100 * mean((model_logit$fitted.values > 0.5) == y)
probit_pcp <- 100 * mean((model_probit$fitted.values > 0.5) == y)
PCP <- c(lmp_pcp, logit_pcp, probit_pcp)
names(PCP) <- c("LMP PCP", "Logit PCP", "Probit PCP")
PCP
```

```
##    LMP PCP  Logit PCP Probit PCP
##   67.43119   68.11927   68.11927
```

It can be seen that the three models are almost equally good at predicting. The logit and probit models only predict marginally better at 68.12%, while the LPM correctly predicts 67.43%.