Exam 2

Mathias Kold

2024-04-18

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

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.

```r
library(readr); library(texreg); library(car)
```

```
## Version:  1.39.3
## Date:     2023-11-09
## Author:   Philip Leifeld (University of Essex)
##
## Consider submitting praise using the praise or praise_interactive functions.
## Please cite the JSS article in your publications -- see citation("texreg").
```

```
## Loading required package: carData
```

```r
data2 <- read_csv("data2.csv")
```

```
## Rows: 450 Columns: 10
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (10): obs, idnumber, salary, lsalary, educ, salbegin, lsalbegin, male, m...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

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.

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

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

```r
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 5.22e-09 ***
## salbegin     1.60816    0.06408  25.097  < 2e-16 ***
## 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: < 2.2e-16
```

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 1.83088 units (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 1.72539 units (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  < 2e-16 ***
## educ           0.023578   0.003993   5.905 7.01e-09 ***
## log(salbegin)  0.820725   0.037051  22.151  < 2e-16 ***
```

```
## 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: < 2.2e-16
```

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.

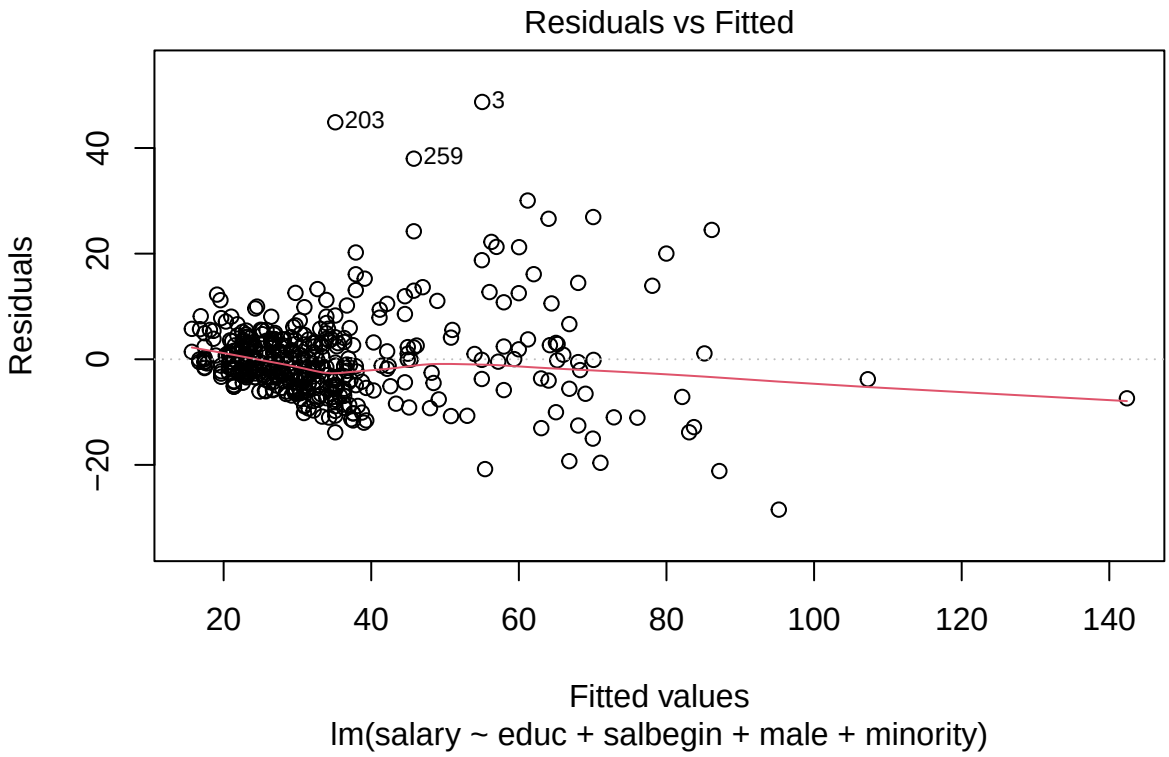
The intercept is 0.84913 which is the value of the dependent variable, in this case annual salary (in 1000 dollars), when all other variables have a value of 0.

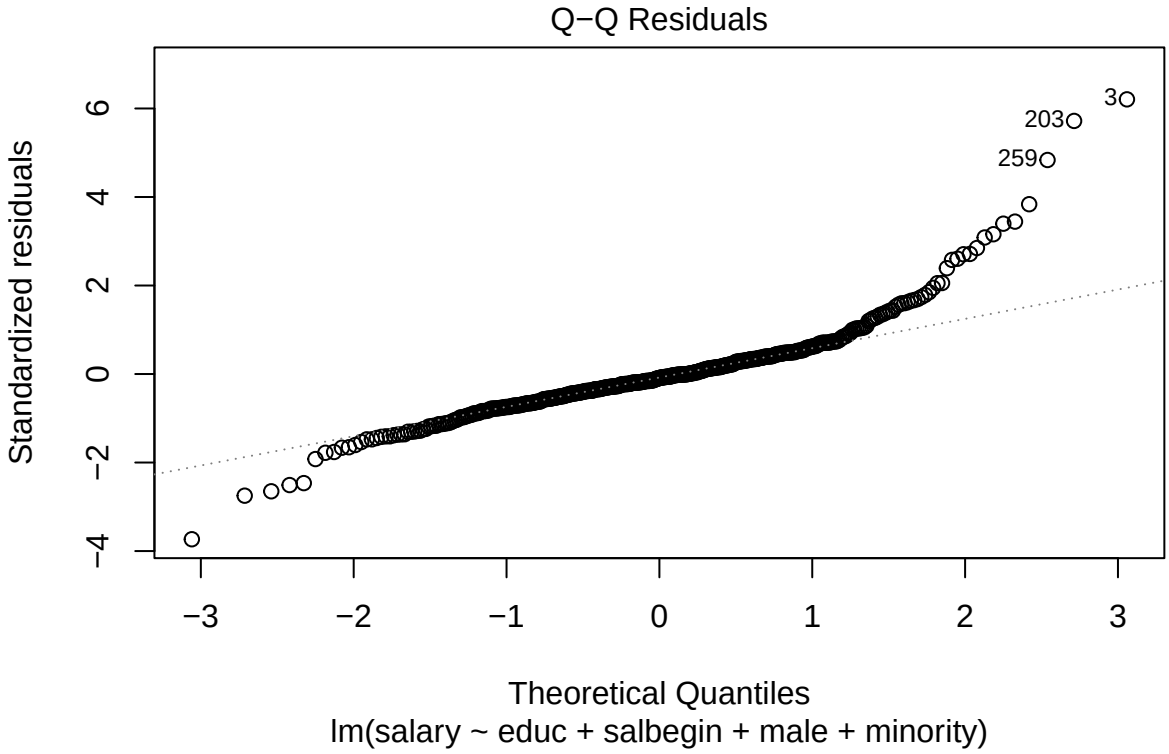
It can also be seen that all the variables are statistically significant at a 5% significance level due to p-values < 0.05. Only education and log(salbegin) are significant at a 1% significance level.

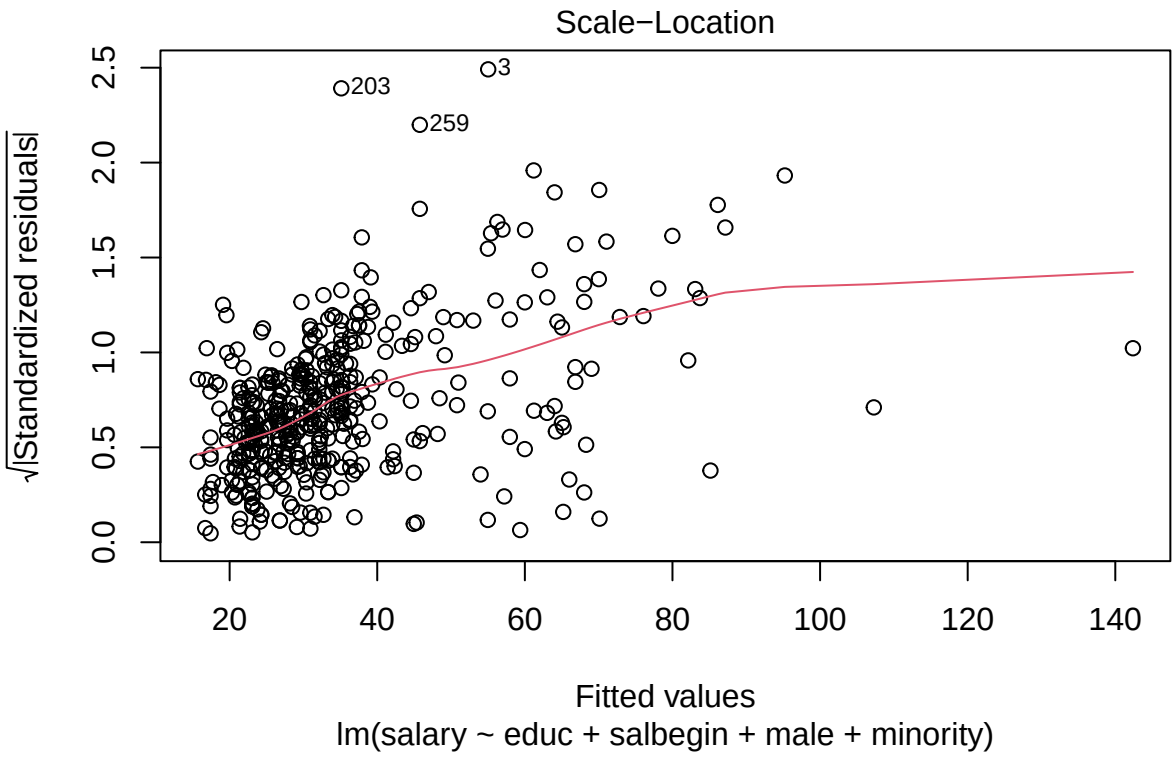
It is worth noting that the adjusted R-squared is 0.8034 for the second model, where as it is a bit lower in the first model at 0.7944.

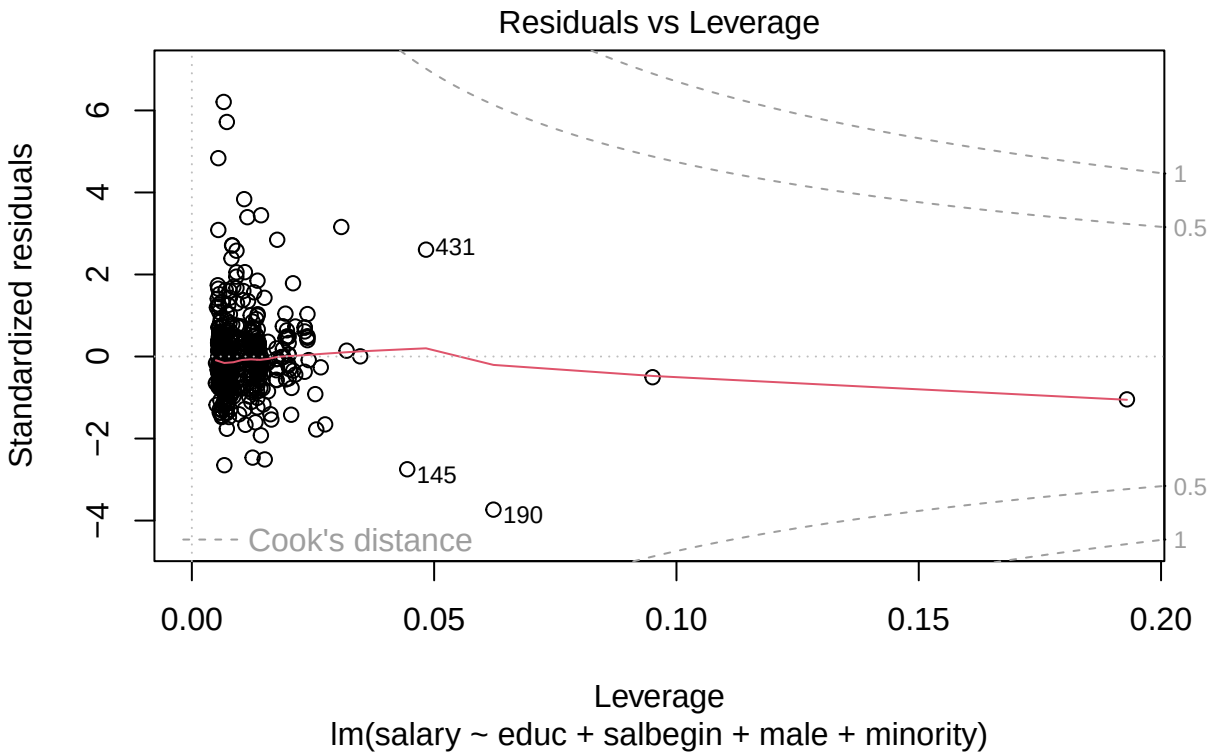
2. Carry out graphical model checking of the two models. Which model would you prefer?

```
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(salary ~ educ + salbegin + male + minority)
3
203
259
−20
0
20
40
20
40
60
80
100
120
140


Q−Q Residuals
Standardized residuals
6
4
2
0
−2
−4
3
203
259
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(salary ~ educ + salbegin + male + minority)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
203
3
259
20
40
60
80
100
120
140
Fitted values
lm(salary ~ educ + salbegin + male + minority)



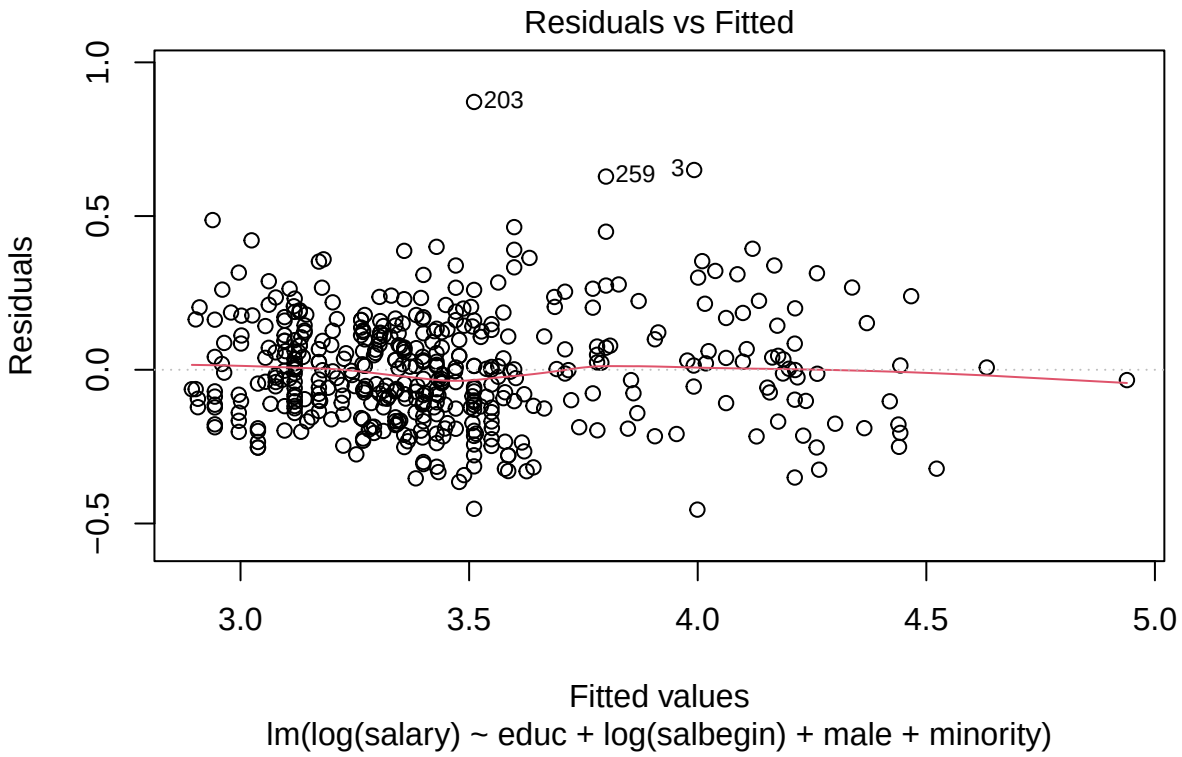
The residual vs fitted model shows if the residuals have non-linear patterns. If they are equally spread around a horizontal line without any patterns it indicates that the residuals does not have non-linear patterns. It can be seen that the residuals are not quite spread around a horizontal line but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

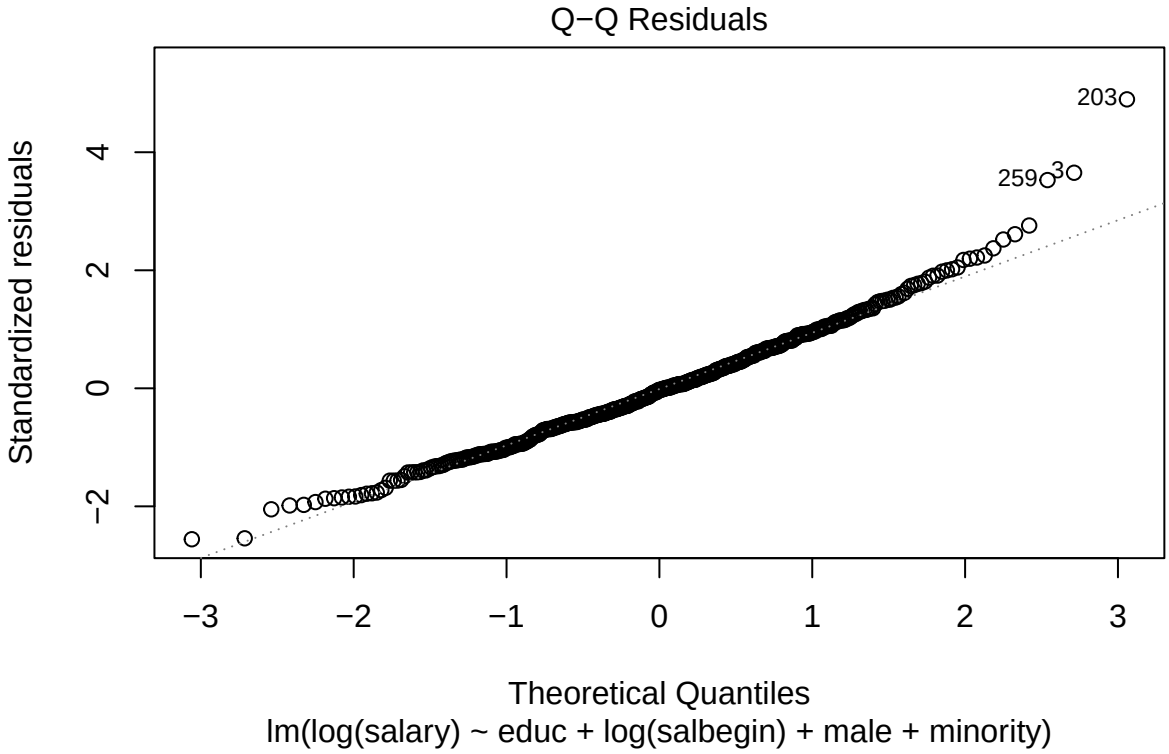
The Q-Q residuals plot shows if the residuals are normally distributed. If the residuals fit the dotted line they are normally distributed. In this case, the residuals fit the dotted line except in the first and last quantiles where they deviate from the dotted line. They deviate a bit more upwards from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

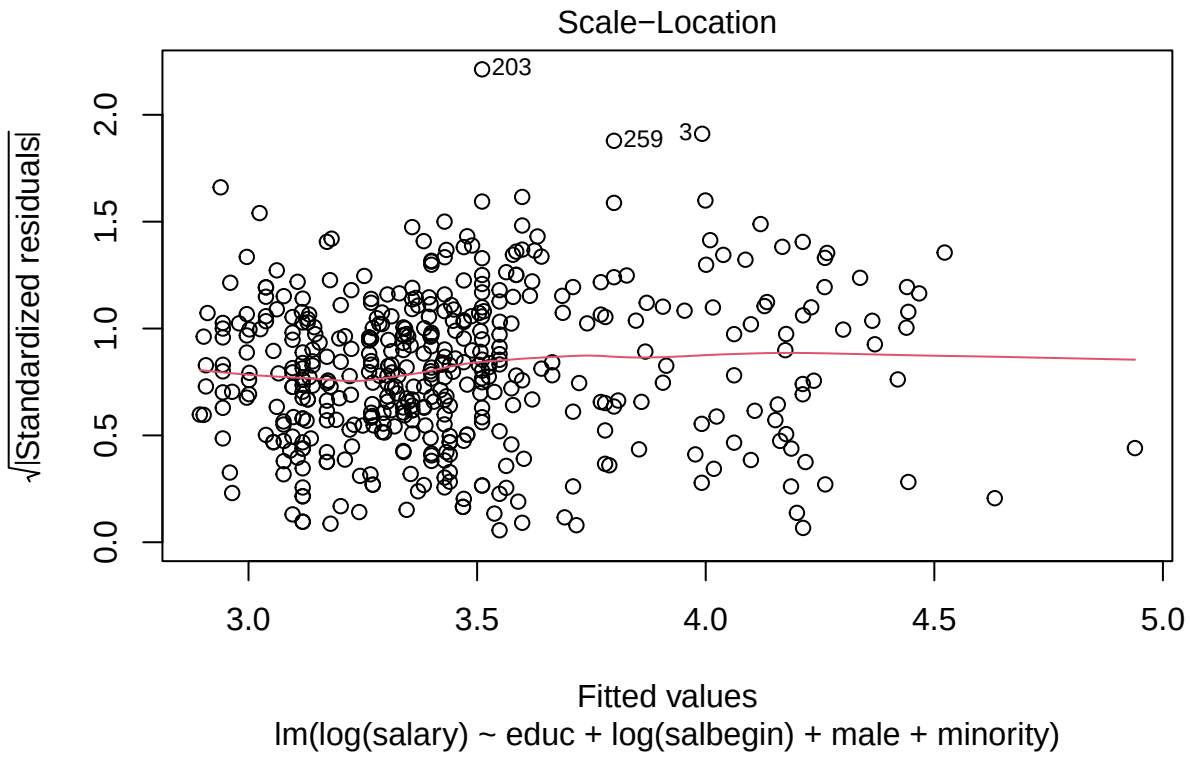
The scale-location plot shows if homoskedasticity exists within the residuals. If the residuals are spread equally around the predictors, the model fulfills the assumption of homoskedasticity. In this case, it can be argued that there is an increase in the spread of the residuals which indicates heteroskedasticity.

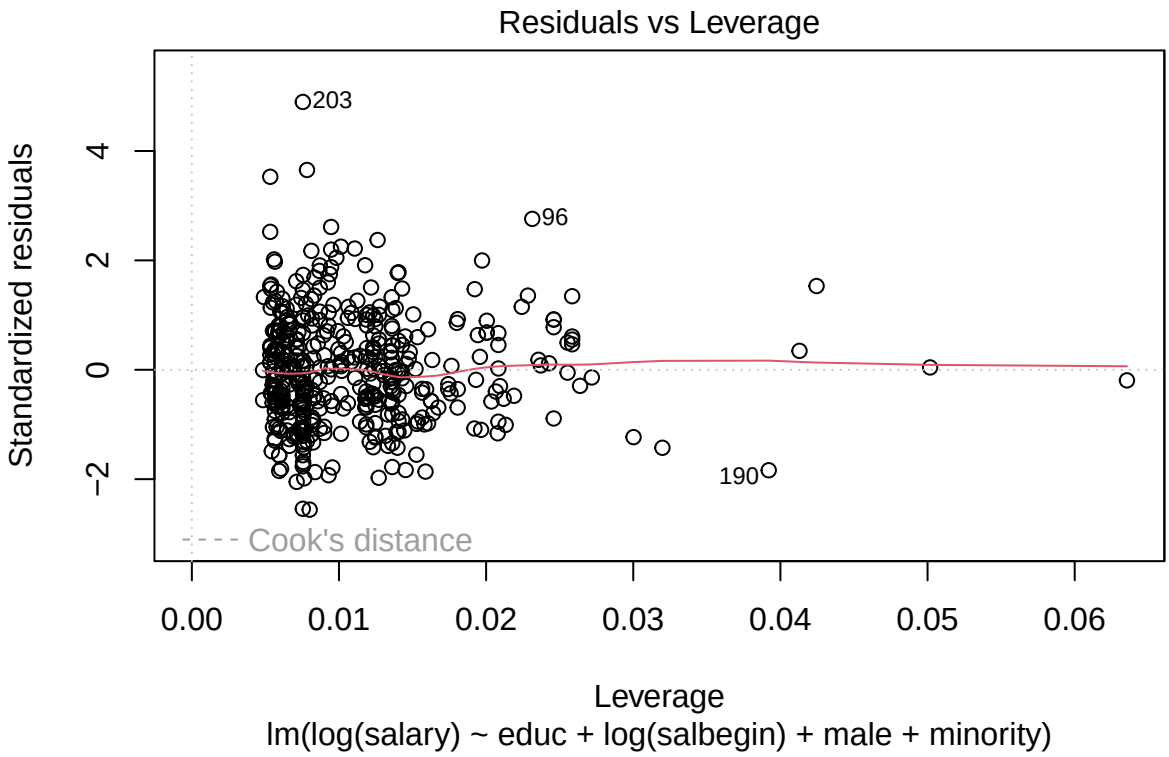
The residuals vs leverage plot helps detect outliers in the model. It is important because outliers can have a major impact in the model. The residuals can be seen if they have a large Cook's distance, which is the dashed line. In this case, it does not seem like there is any observations with a large 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.

```
plot(model2)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
203
259
3
3.0
3.5
4.0
4.5
5.0
Fitted values
lm(log(salary) ~ educ + log(salbegin) + male + minority)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(log(salary) ~ educ + log(salbegin) + male + minority)
203
259
3


Scale−Location
203
259
3
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
3.0
3.5
4.0
4.5
5.0
Fitted values
lm(log(salary) ~ educ + log(salbegin) + male + minority)



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 looks 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, it can be argued that model 2 would be preferred over model 1.

##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:

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

First we use the values found from the two models to find our yhat for model and model2. 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.

```r
yhat1 <- -6.93228 + 0.99327*educ + 1.60816*salbegin + 1.83088*male - 1.72539*minority
yhat2 <- 0.849130 + 0.023578*log(salbegin) + 0.045474*male - -0.041856*minority
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:

```r
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 model2:

```r
library(lmtest)
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)

```r
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 mis-specified. However, both p-values are still pretty low, and would not be accepted at e.g. a 10% significance level.

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. Furthermore, by including $educ^2$, the model will be able to capture any quadratic forms.

```
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  < 2e-16 ***
## 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: < 2.2e-16
```

```
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 2.59e-08 ***
## educ          -0.0147950  0.0229937  -0.643   0.5203
## log(salbegin)  0.7825191  0.0433061  18.069  < 2e-16 ***
```

```
## 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: < 2.2e-16
```

```
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's an upward bias on educ, when $educ^2$ is not included, whereas there's 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.

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 is statistically significant.

The formula for the F-statistics is:

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

where R^2_{ur} is the R-squared from our unrestricted model, R^2_r is the R-squared from the restricted, 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.

«««< HEAD The two unrestricted models are given in exercise 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 unrestriced 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 accepting our alternative hypothesis meaning that β_1 and β_5 for the education does have a statistically 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:

```r
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 1.88e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
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 1.275e-08 ***
## ---
## 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 some answer as before still confirming the rejecting of our null hypothesis.

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

It is of course possible that there could be some kind og 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 taking from tax register, then it might not reflect the truly value of peoples income, because some might have a larger income if they done some kind of undeclared work that is not registered. If the data is obtained by going around and 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 assumption:

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

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 now instead look at a model where we assume a measurement error 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 therefore be 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 instead 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 measurement error 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.