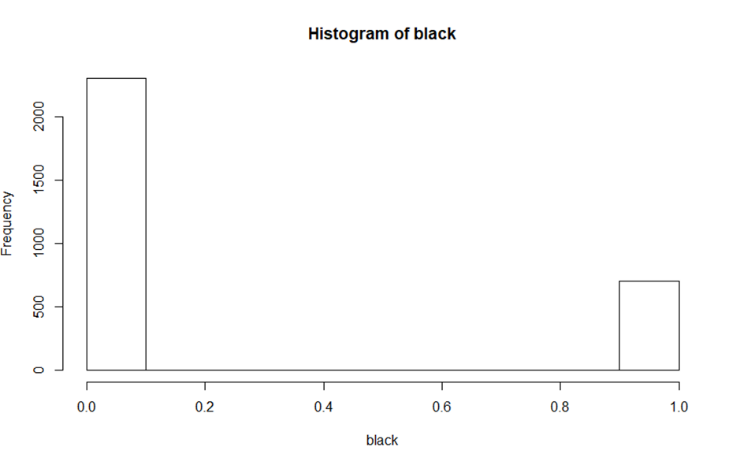
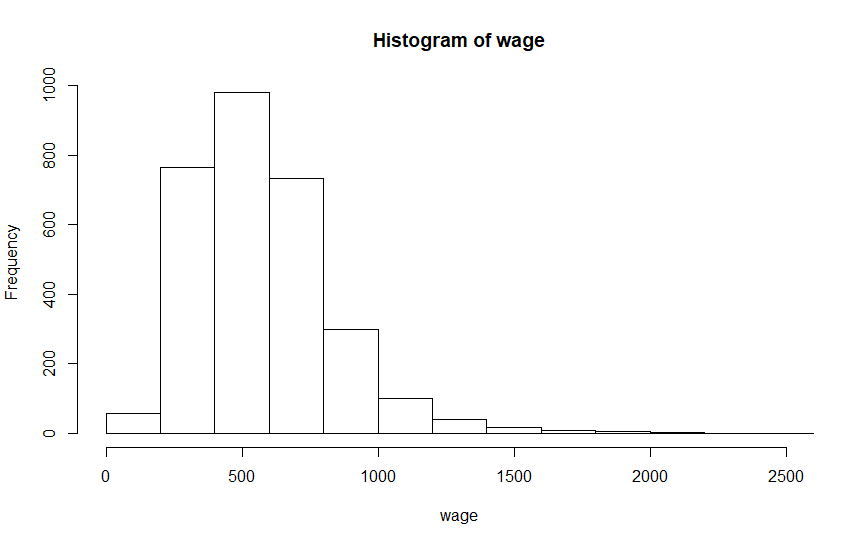
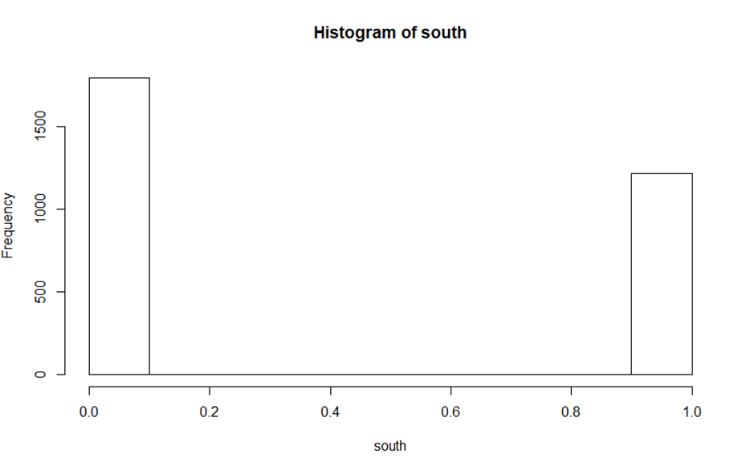
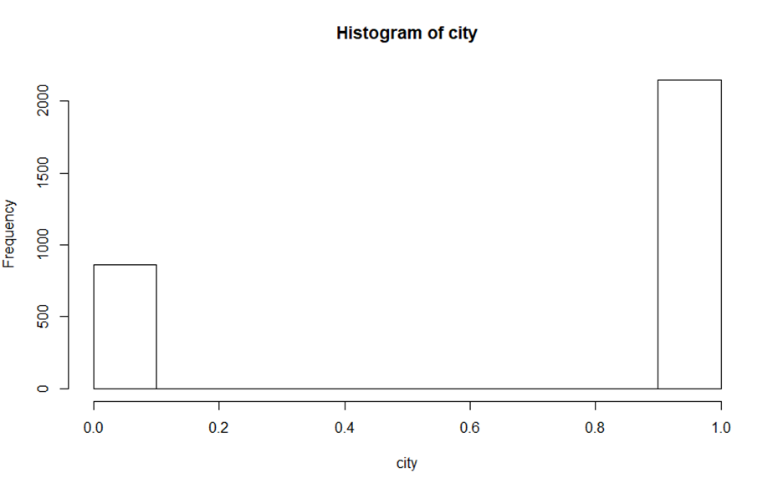
Econometrie

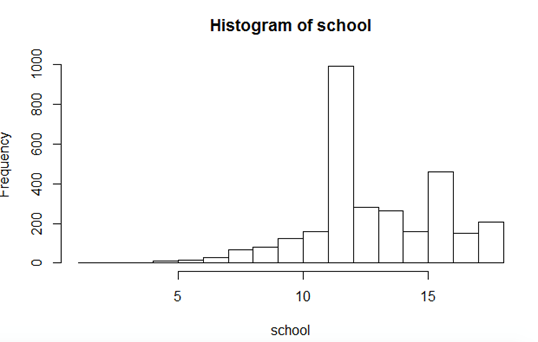
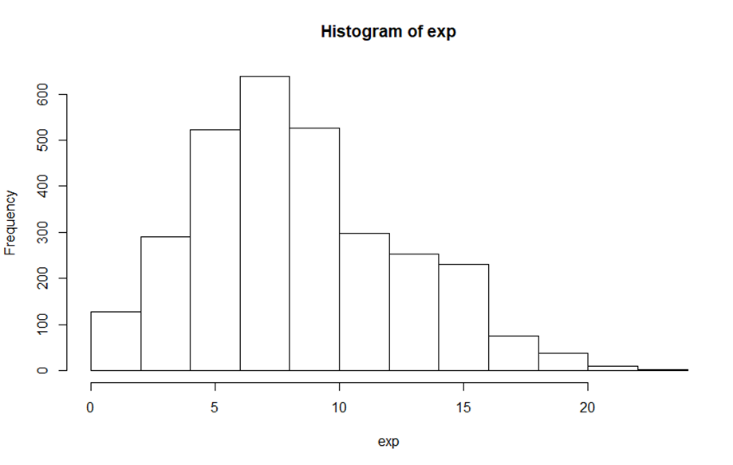
# Step 1: Data description, empirical specification, formulating hypotheses

## Data description

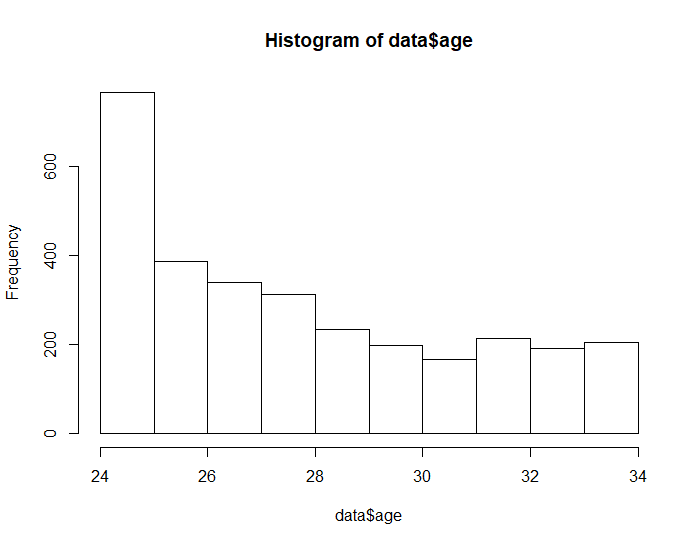


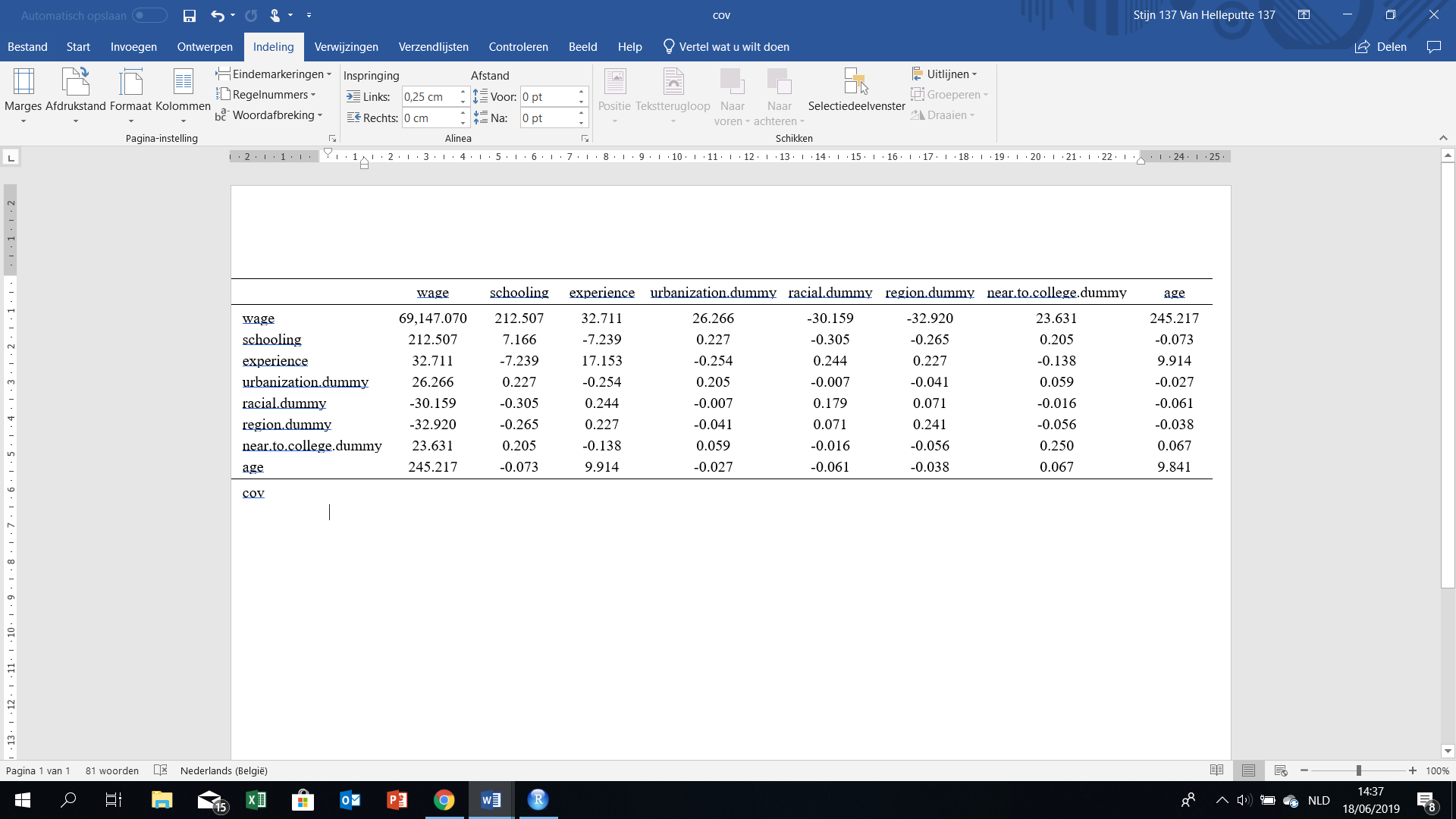


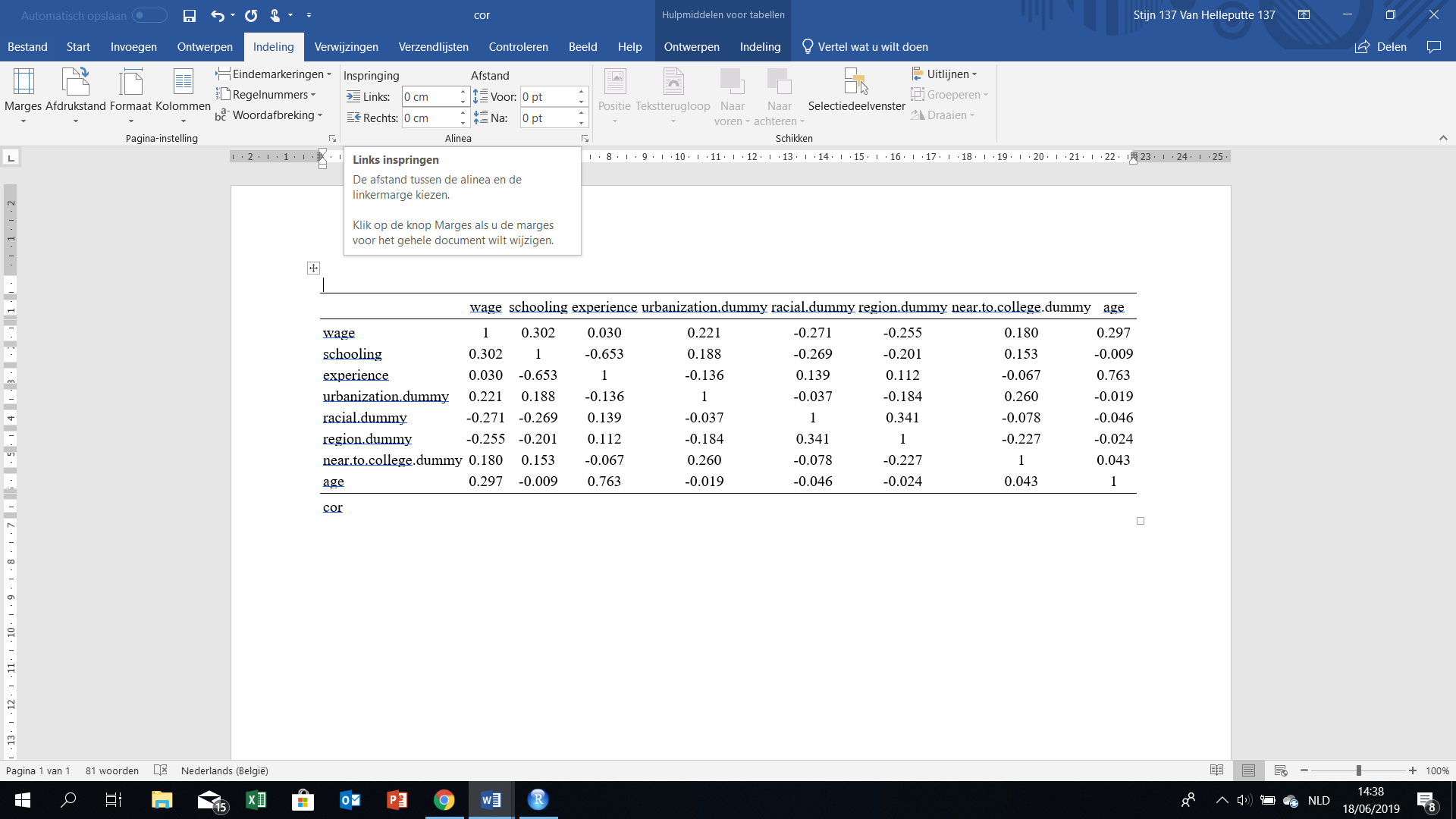


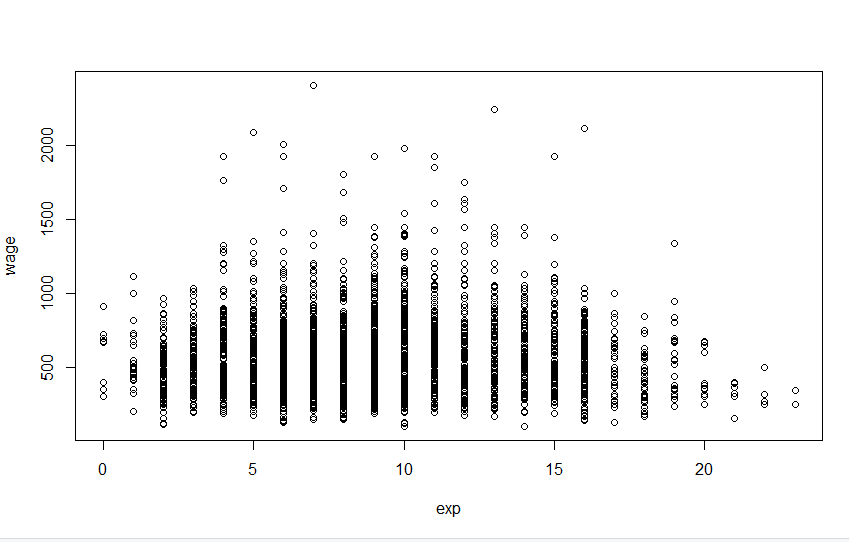
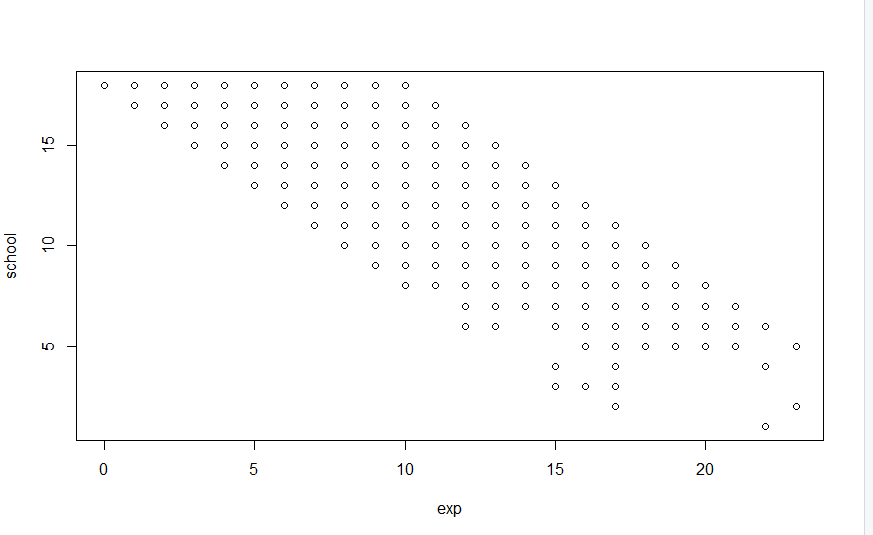


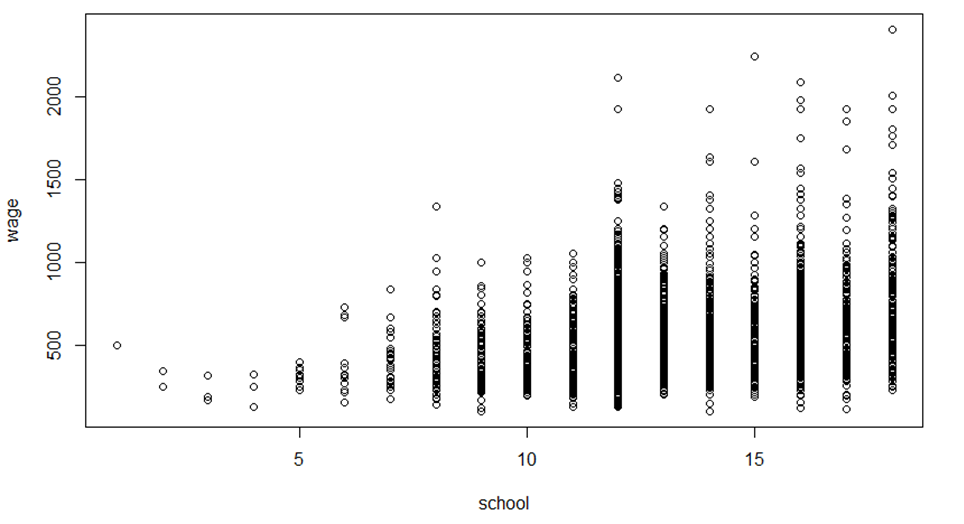
Afbeelding met schermafbeelding

Beschrijving is gegenereerd met zeer hoge betrouwbaarheid

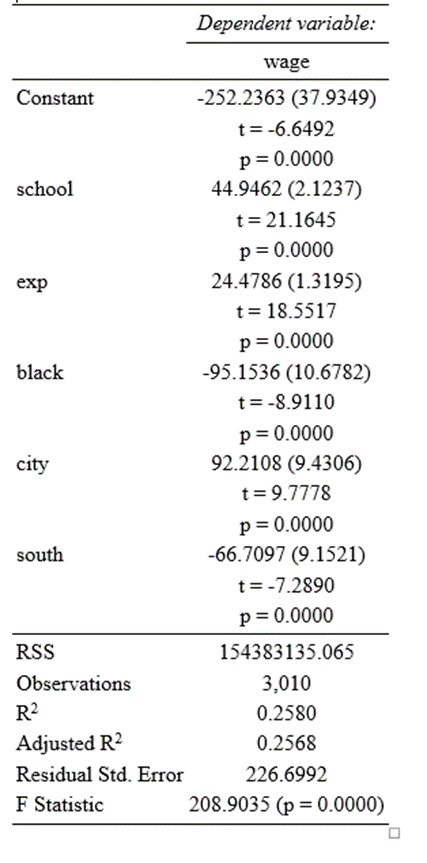








## Empirical specification



## Formulating hypotheses

We chose one-sided hypotheses in all cases, since we already have a clear idea in

what side the decision variable will be influenced. A one-sided test is also less

conservative.

**Experience:**

H0: β2 <= 0

H1: β2 > 0

The proposed hypothesis H0 is that experience has no influence on the wage. The alternative hypothesis H1 is that experience has a positive influence on the wage.

**School:**

H0: β3 <= 0

H1: β3 > 0

The proposed hypothesis H0 is that school has no influence on the wage. The alternative hypothesis H1 is that school has a positive influence on the wage.

**South:**

H0: β4>= 0

H1: β4 < 0

The proposed hypothesis H0 is that south has no influence on the wage. The alternative hypothesis H1 is that south has a negative influence on the wage.

**City:**

H0: β5 <= 0

H1: β5 > 0

The proposed hypothesis H0 is that city has no influence on the wage. The alternative hypothesis H1 is that city has a positive influence on the wage.

**Black:**

H0: β6 >= 0

H1: β6< 0

The proposed hypothesis H0 is that black has no influence on the wage. The alternative hypothesis H1 is that black has a negative influence on the wage.

# Step 2: OLS regression and hypothesis testing

## The model



## Scaling

Experience: als experience stijgt met 1 jaar, dan stijgt je loon met 24.4786

School: als school stijgt met 1 jaar, dan stijgt je loon met 44.9462

South: als je in south woont, dan daalt je loon met 66.7097 (standaard region is other)

City: als je in city woont, dan stijgt je loon met 92.2108 (standaard is countryside)

Black: als je ‘black’ bent, dan daalt je loon met 95.1536 (standaard is white)

## T-testen & F-test

### T

Alle nulhypothesen worden verworpen, elke variabele hierboven besproken heeft in dit model individueel een significante invloed op het loon.

### F

Er is minstens 1 van de parameters β2 , β3 , β4 ,β5 , β6 significant verschillend van 0. Bijgevolg heeft minstens 1 van de verklarende variabelen een significante invloed op het loon.

## Explanatory power

R^2 =0.258

H0: R²=0

H1: R² ≠ 0

p = 0.0000 🡺 We verwerpen onze nulhypothese op 5% significantieniveau

Conclusie:

De “explanatory power” van het model is significant verschillend van 0. Het model heeft dus in zekere mate verklarende waarde/kracht.

## Normality of residuals

H0 : Residuals normaal verdeeld (Residuals ~N)

H1 : Residuals niet normaal verdeeld (Residuals≁N)

Jarque-Bera Normality Test

data: model$residuals

JB = 3687.7, p-value < 2.2e-16

alternative hypothesis: greater

p=0.0000 🡺 We verwerpen onze nulhypothese op 5% significantieniveau

Conclusie: De storingstermen zijn niet normaal verdeeld.

# Step 3: Test the Gauss-Markov assumptions and perform individual remedial measures

## Assumption 1: The regression model is linear in parameters

Wage = β1 + β2 \*exp + β3 \*school + β4 \*south + β5 \*city + β6 \*black + μ

Baseline specification bestaat uit de variabelen experience, school, south, city, black. Deze variabelen worden vermenigvuldigd met hun respectievelijke parameters. Er komen geen machten, logaritmen, sinussen… voor in deze parameters β. Het model is bijgevolg lineair in de parameters.

=> assumption met

## Assumption 2: X is stochastic, but independent of μ

=> assumption met

## Assumption 3: The mean of residuals is zero

=> assumption met

## Assumption 4: Homoscedasticity: The variance of the error terms is constant over time

**Model used:**

lm(formula = wage ~ schooling + experience + urbanization.dummy +

racial.dummy + region.dummy, data = data)

Residuals:

Min 1Q Median 3Q Max

-552.60 -142.64 -25.89 107.22 1583.64

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -252.236 37.935 -6.649 3.49e-11 \*\*\*

schooling 44.946 2.124 21.165 < 2e-16 \*\*\*

experience 24.479 1.319 18.552 < 2e-16 \*\*\*

urbanization.dummy 92.211 9.431 9.778 < 2e-16 \*\*\*

racial.dummy -95.154 10.678 -8.911 < 2e-16 \*\*\*

region.dummy -66.710 9.152 -7.289 3.97e-13 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 226.7 on 3004 degrees of freedom

Multiple R-squared: 0.258, Adjusted R-squared: 0.2568

F-statistic: 208.9 on 5 and 3004 DF, p-value: < 2.2e-16

**Detection**

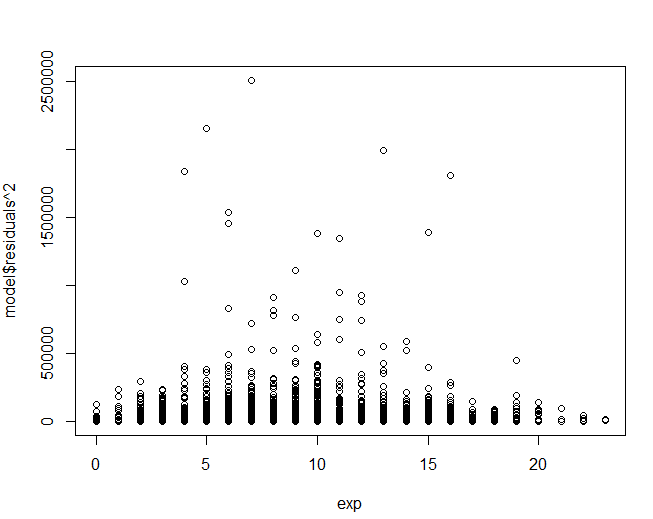
Informal

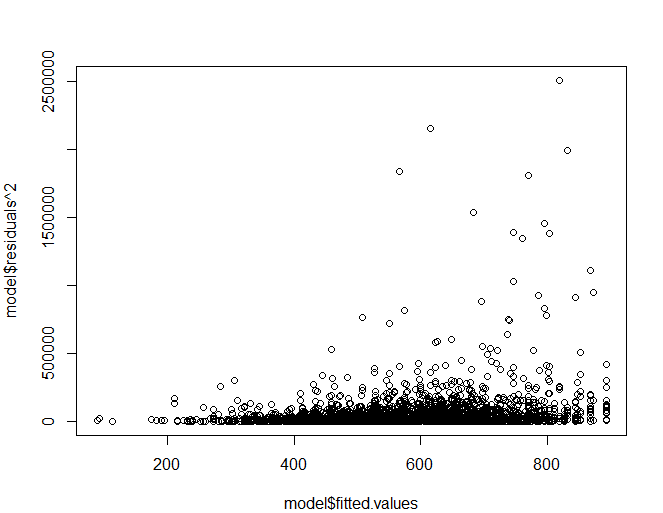
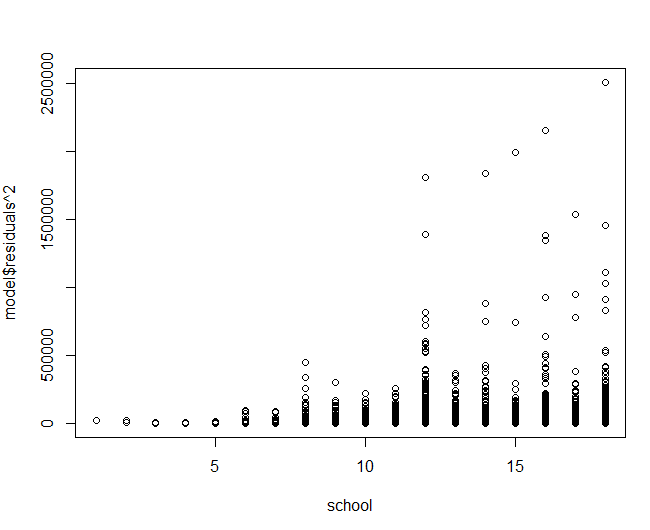
Afbeelding met schermafbeelding

Beschrijving is gegenereerd met zeer hoge betrouwbaarheidAfbeelding met schermafbeelding

Beschrijving is gegenereerd met hoge betrouwbaarheid

Afbeelding met schermafbeelding

Beschrijving is gegenereerd met zeer hoge betrouwbaarheid



Formal

**Goldfeld-Quandt test**

Ass: -σ^2 i is positively related to one of the explanatory variables

-Errorterms are normaly distributed => VIOLATED!!

*Ordered by exp:*

> gqtest(model, fraction = 10,alternative = "two.sided", order.by=exp)

Goldfeld-Quandt test

data: model

GQ = 1.0971, df1 = 1494, df2 = 1494, p-value = 0.07352

alternative hypothesis: variance changes from segment 1 to 2

> gqtest(model, fraction = 10,alternative = "greater", order.by=exp)

Goldfeld-Quandt test

data: model

GQ = 1.0971, df1 = 1494, df2 = 1494, p-value = 0.03676

alternative hypothesis: variance increases from segment 1 to 2

> gqtest(model, fraction = 10,alternative = "less", order.by=exp)

Goldfeld-Quandt test

data: model

GQ = 1.0971, df1 = 1494, df2 = 1494, p-value = 0.9632

alternative hypothesis: variance decreases from segment 1 to 2

*Ordered by school:*

> gqtest(model, fraction = 10,alternative = "two.sided", order.by=school)

Goldfeld-Quandt test

data: model

GQ = 1.5807, df1 = 1494, df2 = 1494, p-value < 2.2e-16

alternative hypothesis: variance changes from segment 1 to 2

> gqtest(model, fraction = 10,alternative = "greater", order.by=school)

Goldfeld-Quandt test

data: model

GQ = 1.5807, df1 = 1494, df2 = 1494, p-value < 2.2e-16

alternative hypothesis: variance increases from segment 1 to 2

> gqtest(model, fraction = 10,alternative = "less", order.by=school)

Goldfeld-Quandt test

data: model

GQ = 1.5807, df1 = 1494, df2 = 1494, p-value = 1

alternative hypothesis: variance decreases from segment 1 to 2

If we order according to experience, the variance increases. If we order according to school, the variance increases. We reject the null hypothesis and assume heteroskedasticity

**White’s General Heteroskedasticity test**

##Assumption: non

> expXschool <- exp\*school

> expXblack <- exp\*black

> expXsouth <- exp\*south

> expXcity <- exp\*city

> schoolXblack <- school\*black

> schoolXsouth <- school\*south

> schoolXcity <- school\*city

> blackXsouth <- black\*south

> blackXcity <- black\*city

> southXcity <- south\*city

Met kruisproduct

> bptest(model, ~ school+I(school^2)+exp+I(exp^2)+black+I(black^2)+south+I(south^2)+city+I(city^2)+expXschool+expXblack+expXsouth+expXcity+schoolXblack+schoolXsouth+schoolXcity+blackXsouth+blackXcity+southXcity)

studentized Breusch-Pagan test

data: model

BP = 154.25, df = 17, p-value < 2.2e-16

P value smaller than 0.05, so we reject the null hypothesis. When calculated manually we get the same results. Hetrosk + autocorrelation

Zonder kruisproduct

> bptest(model, ~ school+I(school^2)+exp+I(exp^2)+black+I(black^2)+south+I(south^2)+city+I(city^2))

studentized Breusch-Pagan test

data: model

BP = 125.83, df = 7, p-value < 2.2e-16

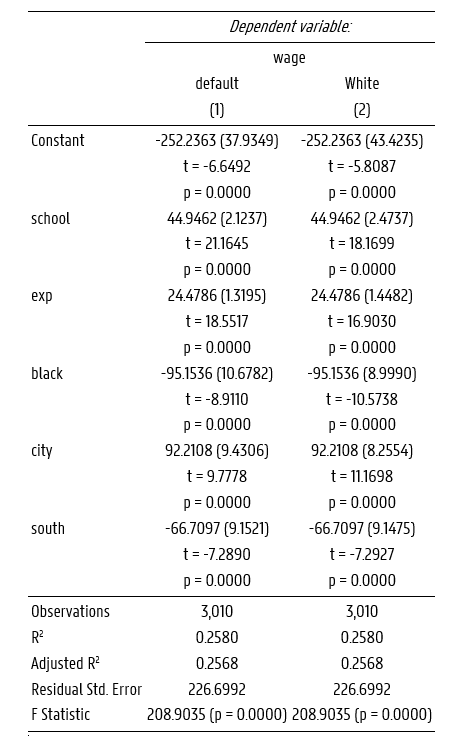
P value smaller than 0.05, so we reject the null hypothesis. When calculated manually we get the same results. Pure hetrosk.

**Redemedial Measures**

is not known (and can not be estimated consistently)

Aangezien we voor elke x waarde (een combinatie van city, black, south, schooling en experience) maar 1 y-waarde hebben, kunnen we hier geen schatten voor elke x-waarde. EGLS en GLS zijn hier niet toepasbaar.

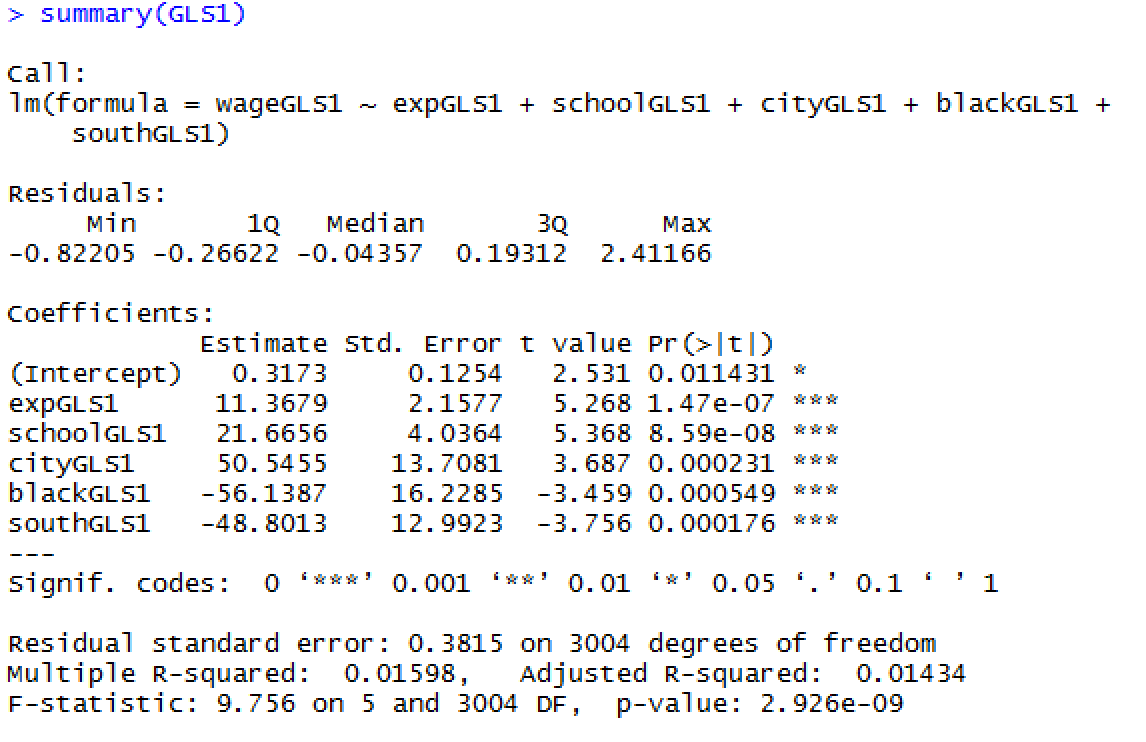
**White’s heteroskedasticity consistent variance**

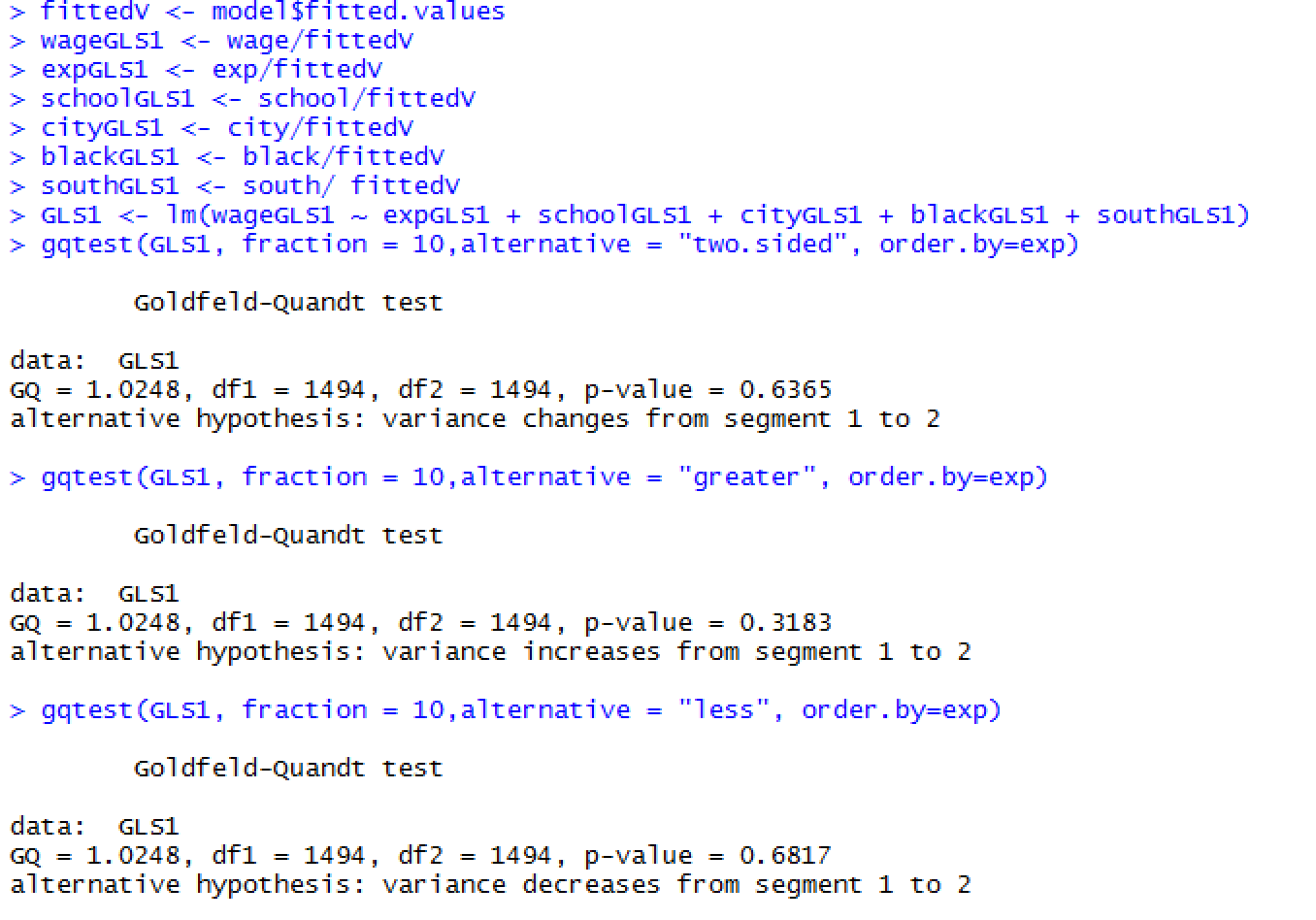


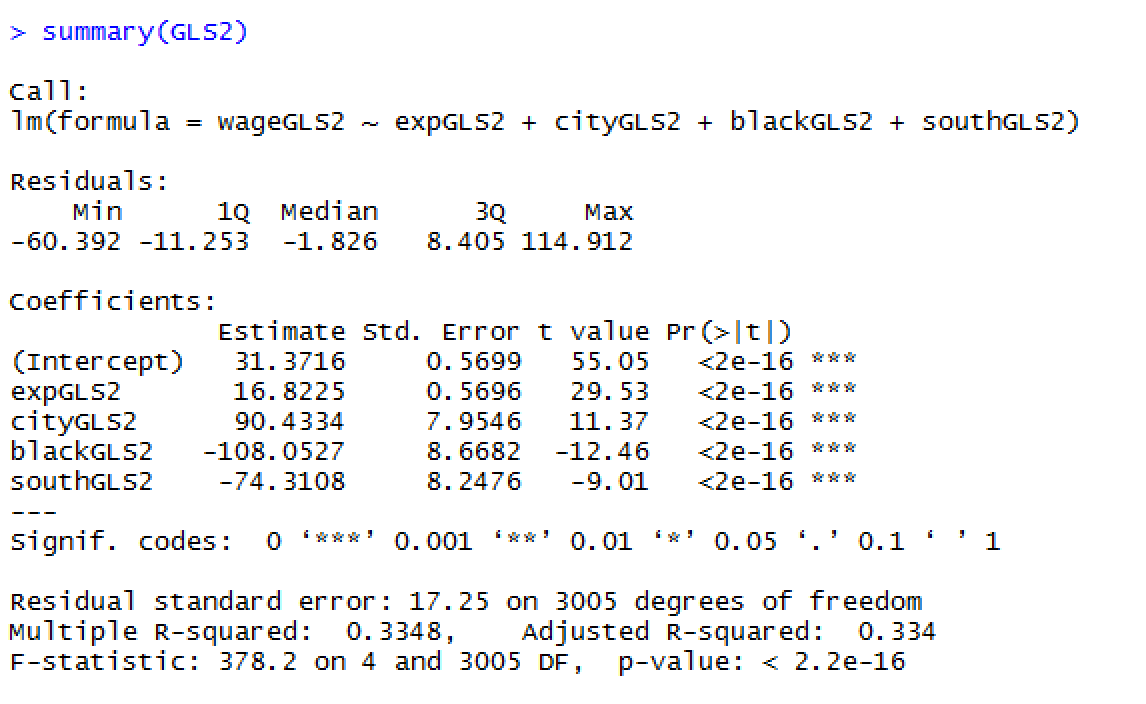
**Ad-hoc transforms - GLS**

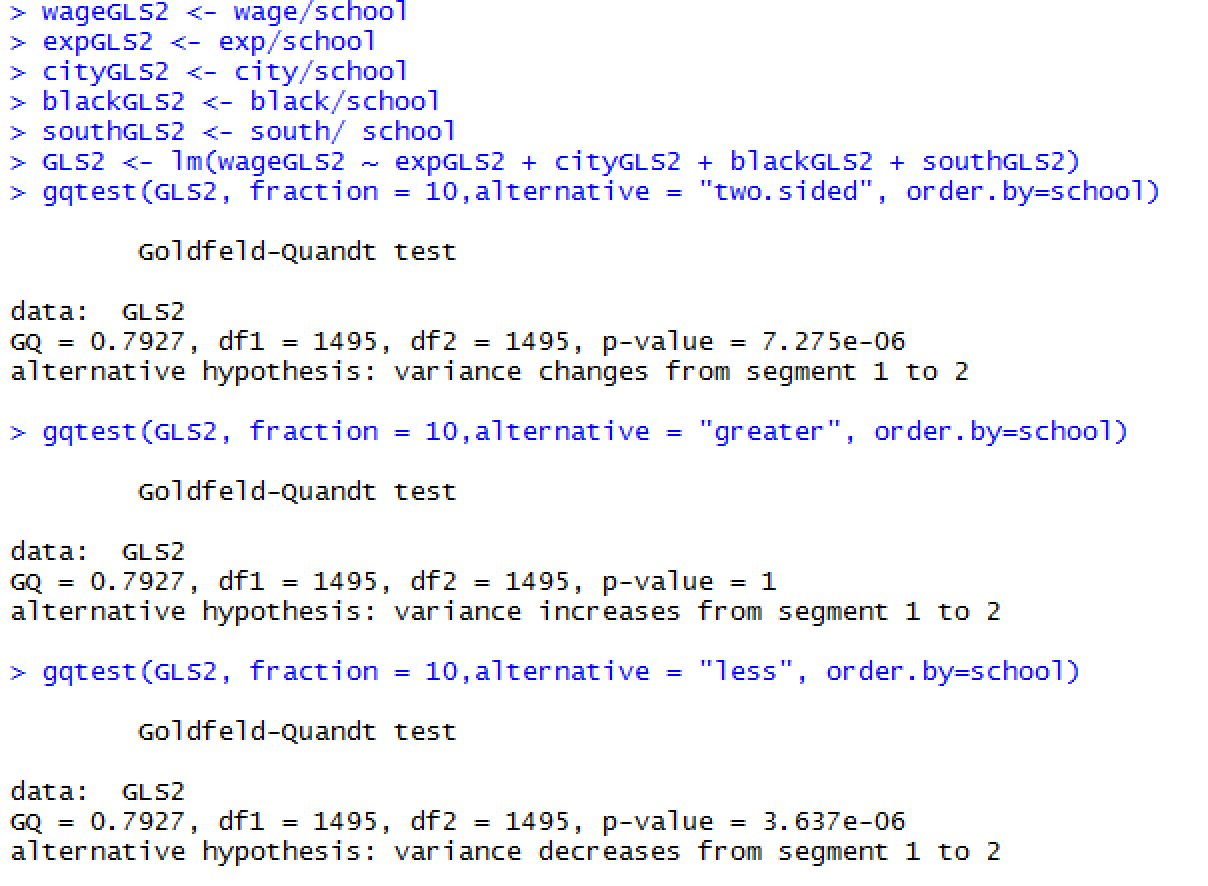
3 mogelijke transformaties -> assumpties omtrent dummies en experience (0 waarden) niet relevant

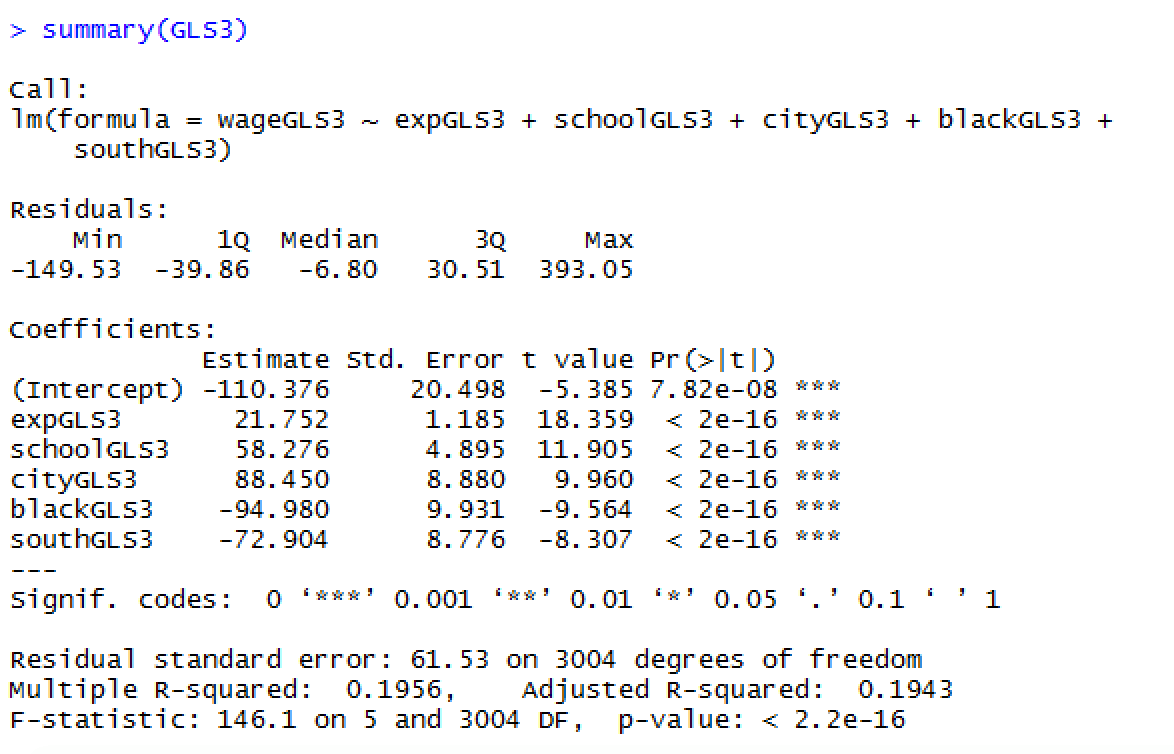
1. Delen door fitted values
2. Delen door schooling
3. Delen door wortel schooling

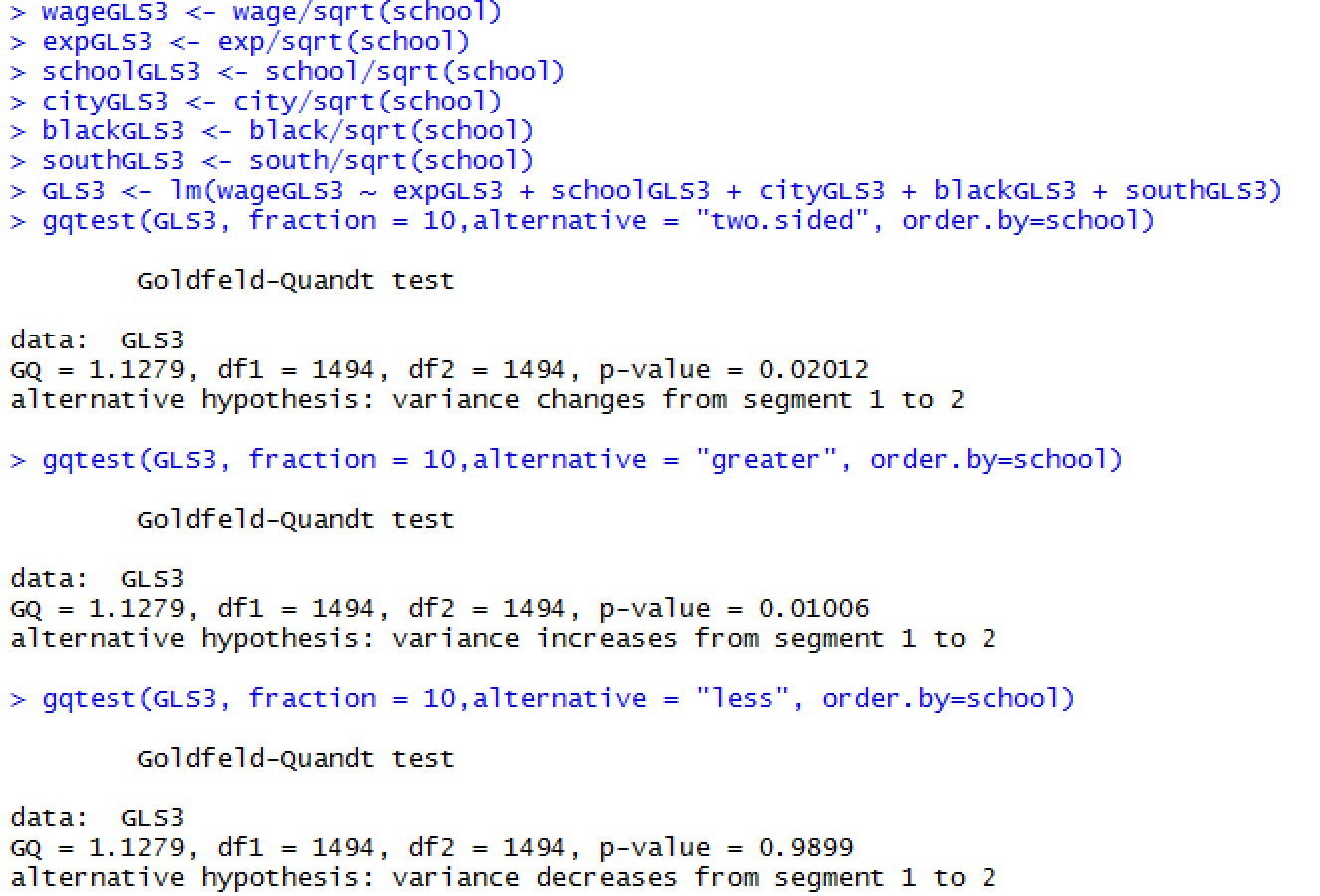












Specification/data issues rather than ‘pure’ heteroskedasticity

**Log transform**

Call:

lm(formula = log(wage) ~ log(schooling) + experience + urbanization.dummy +

racial.dummy + region.dummy, data = data)

Residuals:

Min 1Q Median 3Q Max

-1.65265 -0.22912 0.01731 0.24850 1.57184

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.76360 0.11977 31.422 < 2e-16 \*\*\*

log(schooling) 0.83487 0.04118 20.273 < 2e-16 \*\*\*

experience 0.03811 0.00219 17.406 < 2e-16 \*\*\*

urbanization.dummy 0.16598 0.01575 10.539 < 2e-16 \*\*\*

racial.dummy -0.19543 0.01778 -10.990 < 2e-16 \*\*\*

region.dummy -0.12526 0.01530 -8.185 3.97e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3786 on 3004 degrees of freedom

Multiple R-squared: 0.2734, Adjusted R-squared: 0.2722

F-statistic: 226.1 on 5 and 3004 DF, p-value: < 2.2e-16

> gqtest(logModel, fraction = 10,alternative = "two.sided", order.by=exp)

Goldfeld-Quandt test

data: logModel

GQ = 0.94677, df1 = 1494, df2 = 1494, p-value = 0.2905

alternative hypothesis: variance changes from segment 1 to 2

> gqtest(logModel, fraction = 10,alternative = "greater", order.by=exp)

Goldfeld-Quandt test

data: logModel

GQ = 0.94677, df1 = 1494, df2 = 1494, p-value = 0.8547

alternative hypothesis: variance increases from segment 1 to 2

> gqtest(logModel, fraction = 10,alternative = "less", order.by=exp)

Goldfeld-Quandt test

data: logModel

GQ = 0.94677, df1 = 1494, df2 = 1494, p-value = 0.1453

alternative hypothesis: variance decreases from segment 1 to 2

> gqtest(logModel, fraction = 10,alternative = "two.sided", order.by=school)

Goldfeld-Quandt test

data: logModel

GQ = 1.0413, df1 = 1494, df2 = 1494, p-value = 0.4346

alternative hypothesis: variance changes from segment 1 to 2

> gqtest(logModel, fraction = 10,alternative = "greater", order.by=school)

Goldfeld-Quandt test

data: logModel

GQ = 1.0413, df1 = 1494, df2 = 1494, p-value = 0.2173

alternative hypothesis: variance increases from segment 1 to 2

> gqtest(logModel, fraction = 10,alternative = "less", order.by=school)

Goldfeld-Quandt test

data: logModel

GQ = 1.0413, df1 = 1494, df2 = 1494, p-value = 0.7827

alternative hypothesis: variance decreases from segment 1 to 2

## Assumption 5: Checking for autocorrelation

**Detection**

Grafisch:



Aan de hand van de eerste test zien we geen indicaties van een systematisch patroon. Dit ondersteunt de non-autocorrelatie assumptie van het CLRM.

Testen

**Runs test**

> runs(model)

Observed Runs N1 N2

1402 1349 1661





> E-(1.96\*sqrt(V))

[1] 1436.65

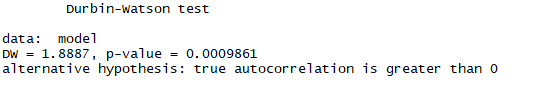
> E+(1.96\*sqrt(V))

[1] 1543.009

Onze R= 1402 valt niet binnen het interval [1436,65:1543,01]

**Durbin-Watson test**

> dwtest(model)



onze d statistiek ligt dicht bij 2 en als we rho hoed bereken bekomen we

https://lh4.googleusercontent.com/YzD32kBV-yaiewiRPz1V2bdXEShXu3tA4JuKXRYN5WbbBtmvAHKkn9xW_VJAyXR6zhCEMK4BCn0hgKvP2HML0xI8RExyFEF5Xi9jrom6we7Vbu9yNsX5Iy-OEpJ3843bfh08DKTl

= 0.05565, wat dicht bij de 0 ligt

Hieruit zouden we al kunnen afleiden dat er geen eerste orde autocorrelatie is.

Maar, p-waarde is echter kleiner dan 0.05 waardoor we de nul hypothese dat er geen autocorrelatie is verwerpen.

Echter is niet voldaan aan enkele van de assumpties nl. dat de verklarende variabelen deterministic zijn en dat de error term normaal verdeeld is.

**Breusch-Godfrey LM test**

> bgtest(model,order = 1)

Breusch-Godfrey test for serial correlation of order up to 1

data: model

LM test = 9.3188, df = 1, p-value = 0.002268

> bgtest(model,order = 2)

Breusch-Godfrey test for serial correlation of order up to 2

data: model

LM test = 10.081, df = 2, p-value = 0.00647

> bgtest(model,order = 3)

Breusch-Godfrey test for serial correlation of order up to 3

data: model

LM test = 11.551, df = 3, p-value = 0.00909

> bgtest(model,order = 4)

Breusch-Godfrey test for serial correlation of order up to 4

data: model

LM test = 12.121, df = 4, p-value = 0.01647

> bgtest(model,order = 5)

Breusch-Godfrey test for serial correlation of order up to 5

data: model

LM test = 14.633, df = 5, p-value = 0.01205

> bgtest(model,order = 6)

Breusch-Godfrey test for serial correlation of order up to 6

data: model

LM test = 14.924, df = 6, p-value = 0.02086

> bgtest(model,order = 7)

Breusch-Godfrey test for serial correlation of order up to 7

data: model

LM test = 16.519, df = 7, p-value = 0.02077

> bgtest(model,order = 8)

Breusch-Godfrey test for serial correlation of order up to 8

data: model

LM test = 16.534, df = 8, p-value = 0.03534

> bgtest(model,order = 9)

Breusch-Godfrey test for serial correlation of order up to 9

data: model

LM test = 21.426, df = 9, p-value = 0.01089

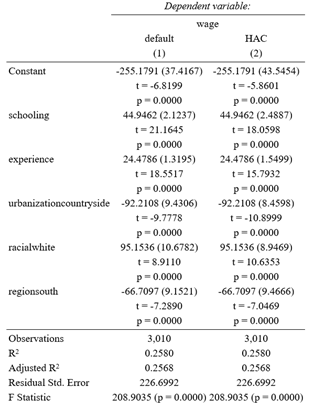
Bij 8 lag values bekomen we de hoogste p waarde, ook hier gelden dezelfde conclusies. H0 wordt verworpen. Er is autocorrelatie.

=> assumption NOT met

**Remedial measures**

Newey-West method of correcting the OLS standard errors

Met deze methode gebruiken we nog steeds OLS, maar we passen de standard errors aan (=HAC standard errors). We zien geen wijzigingen in de coëfficiënten. Deze methode heeft dus geen grote impact als remedial measure, ondanks onze vrij grote sample.



## Assumption 6: The number of observations must be greater than number of parameters

> dim(data)

[1] 3010 13

> model$rank

[1] 6

3010>6

=> assumption met

## Assumption 7: The variability in X values is positive (X values in a given sample must not all be the same)

> var(school)

[1] 7.165862

> var(exp)

[1] 17.15344

> var(south)

[1] 0.2407975

> var(black)

[1] 0.1790665

> var(city)

[1] 0.2047174

=> Assumption met

## Assumption 8: There is no exact collinearity between the X variables -> no perfect multicollinearity

**1. compare R^2 with the t-values**

Theory:

If there is a high R^2, it means that the model has a lot of explanatory power. If the t-values would be low, they would probably not be significant. A combination of high R^2 and low t-values would be an indication of multicollinearity.

We found:

> summary(model)

Call:

lm(formula = wage ~ schooling + experience + urbanization.dummy +

racial.dummy + region.dummy, data = data)

Residuals:

Min 1Q Median 3Q Max

-552.60 -142.64 -25.89 107.22 1583.64

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -252.236 37.935 -6.649 3.49e-11 \*\*\*

schooling 44.946 2.124 21.165 < 2e-16 \*\*\*

experience 24.479 1.319 18.552 < 2e-16 \*\*\*

urbanization.dummy 92.211 9.431 9.778 < 2e-16 \*\*\*

racial.dummy -95.154 10.678 -8.911 < 2e-16 \*\*\*

region.dummy -66.710 9.152 -7.289 3.97e-13 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 226.7 on 3004 degrees of freedom

Multiple R-squared: 0.258, Adjusted R-squared: 0.2568

F-statistic: 208.9 on 5 and 3004 DF, p-value: < 2.2e-16

low R²

high t-values

Conclusion:

There is no reason to assume multicollinearity.

**2. calculate pairwise correlation between explanatory variables**

Theory:

There is high correlation if: |cor| >= 0.8

We found:

|  |
| --- |
| > cor(exp, school)  [1] -0.6529563  > cor(exp, south)  [1] 0.1116499  > cor(exp, city)  [1] -0.1357221  > cor(exp, black)  [1] 0.1392091  > cor(school, south)  [1] -0.2014114  > cor(school, city)  [1] 0.187579  > cor(school, black)  [1] -0.2693878  > cor(south, city)  [1] -0.1844739  > cor(south, black)  [1] 0.3412674  > cor(city, black)  [1] -0.03681341 |
|  |

Conclusion:

None of the correlations are in absolute value higher than 0.8. There is no sign of high multicollinearity. There is however a moderate correlation between experience and school of -0.6. This means that in general, people who went to school longer have less experience, and people who went to school less long have more experience.

**3. calculate auxiliary regressions**

> Reg1<-lm(school~south+black+city+exp)

> Reg2<-lm(exp~school+south+black+city)

> Reg3<-lm(south~school+exp+city+black)

> Reg4<-lm(black~school+exp+city+south)

> Reg5<-lm(city~school+exp+black+south)

> summary(Reg1)$r.squared

[1] 0.4715056

> summary(Reg2)$r.squared

[1] 0.4281006

> summary(Reg3)$r.squared

[1] 0.1531976

> summary(Reg4)$r.squared

[1] 0.1634929

> summary(Reg5)$r.squared

[1] 0.06190826

Conclusion:

The degree to which every X-variable separately can be explained by all other X-variables is measured by the R^2. We find relatively low values. Only for experience and school we find a slightly higher R^2. There is no indication for multicollinearity.

**4. Variance Inflation Factor**

Theory:

If the VIF is larger than 10, it is an indicator for high multicollinearity, if the VIF is larger than 4, it is an indicator for multicollinearity VIF= 1/(1-R²). We use the R² from 3, regressing all variables separately on the other variables. (see above)

> 1/(1-summary(Reg1)$r.squared)

[1] 1.892167

> 1/(1-summary(Reg2)$r.squared)

[1] 1.748559

> 1/(1-summary(Reg3)$r.squared)

[1] 1.180913

> 1/(1-summary(Reg4)$r.squared)

[1] 1.195447

> 1/(1-summary(Reg5)$r.squared)

[1] 1.065994

Conclusion:

The VIF are all very low. None of the values are larger than 4.

There is no indication for multicollinearity.

=> assumption met

## Assumption 9: The regression model is correctly specified

Pattern in the error terms



Not immediately a specific pattern visible.

Heteroskedasticity and/or autocorrelation in the error terms

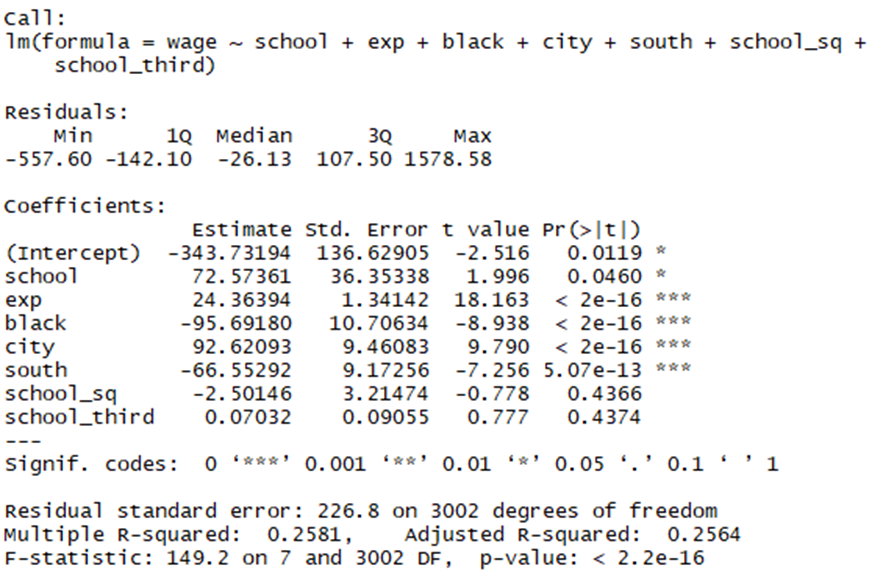
See assumption 4 and 5 above. We have both heteroskedasticity and autocorrelation, these are probably due to a specification error.

Overfit the model

1. Add school^2 en school^3 as parameter

→ both not significant (t-values)

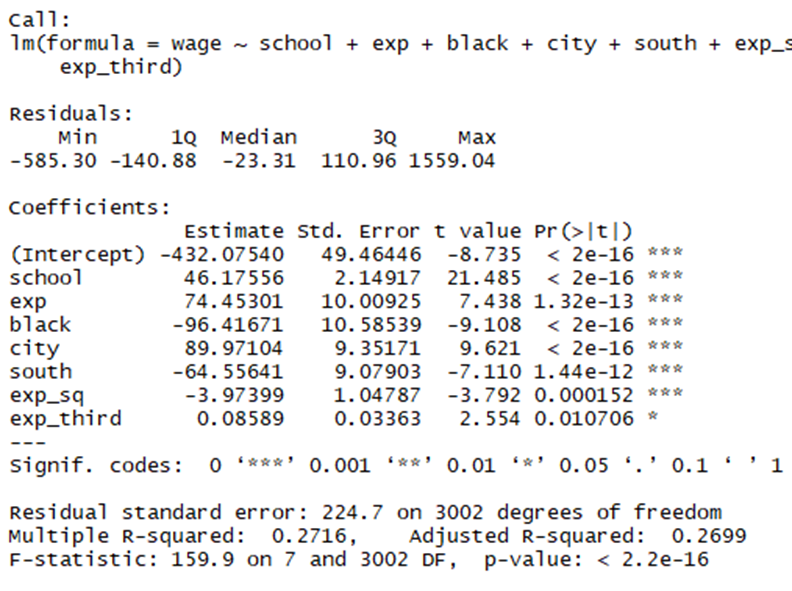
This model results in an r-squared of 0.26. This low R^2 can indicate that there is a presence of more factors in our residuals and this indicates a possible specification error. One possibility is that we need to add another parameter such as the kind of job performed. E.g. a lawyer earns more than a gardener.



1. Add exp^2 en exp^3 as parameter

→ both are significant (t-values)

This model results in an r-squared of 0.27. This low R^2 can indicate that there is a presence of more factors in our residuals like we mentioned above.



*Ramsey’s RESET test*

Onder de nul hypothese zijn alle coefficiënten gelijk aan 0. De p-waarde die we verkrijgen is zeer klein. Daarom verwerpen we de nul hypothese (alle coëfficienten zijn tegelijk 0), in andere woorden : er is een specificatie fout volgens de Ramsey test.



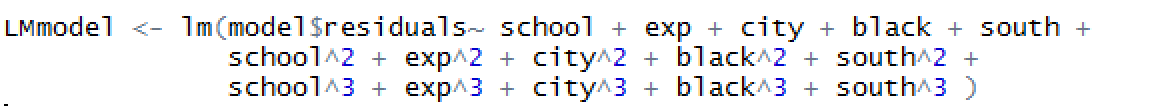
Conclusie: er is een specificatiefout.

*Lagrange Multiplier (LM) test*

Bij deze test voor functional form wordt een lineair model geschat met de geschatte error-terms (van de baseline specification) als dependent variable en de explainatory variables zijn de explainatory variables van de baseline specification en deze variabelen tot de 2e en 3e macht.

Bij een correcte functional form zouden deze variabelen tot de 2e en 3e macht geen invloed mogen hebben op de dependent variable. De bijhorende Bèta’s zouden dus insignificant moeten zijn. Aangezien we geen correcte test vonden op R proberen we zelf de test uit te voeren.

*Model*



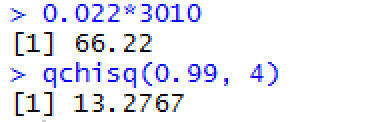
Wanneer de dummy variables tot een macht verhoffen worden blijven de waarden dezelfde dus worden deze bijgevolg uit het model gehaald.



*Summary*



R2\*n = 0.022\*3010:

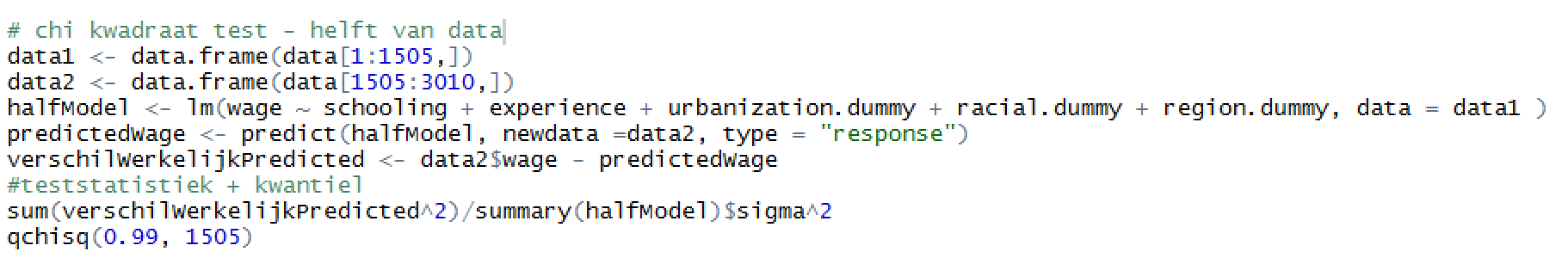


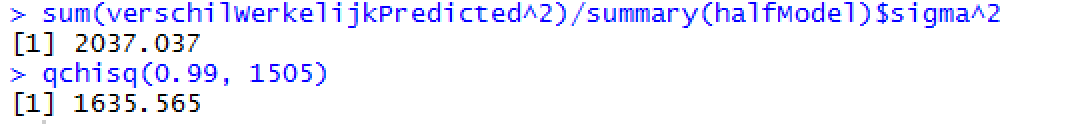
Conclusie: er is een specificatiefout.

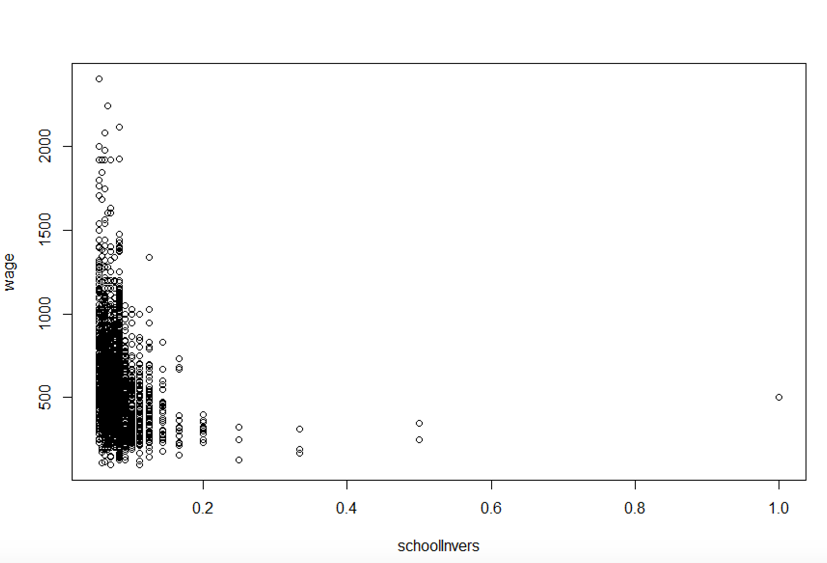
*Forecast chi squared test*

Requires : a sufficiently large sample --> n=3010

N1 = 1505

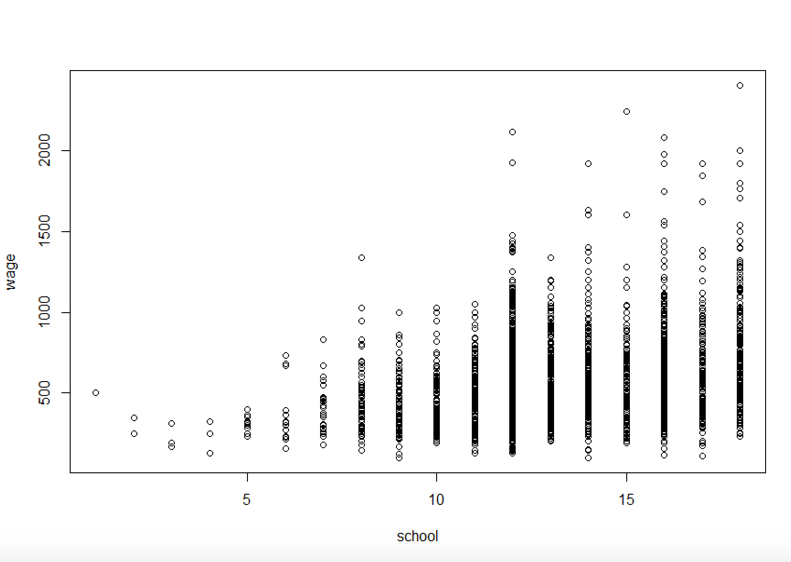
N2 = 1505

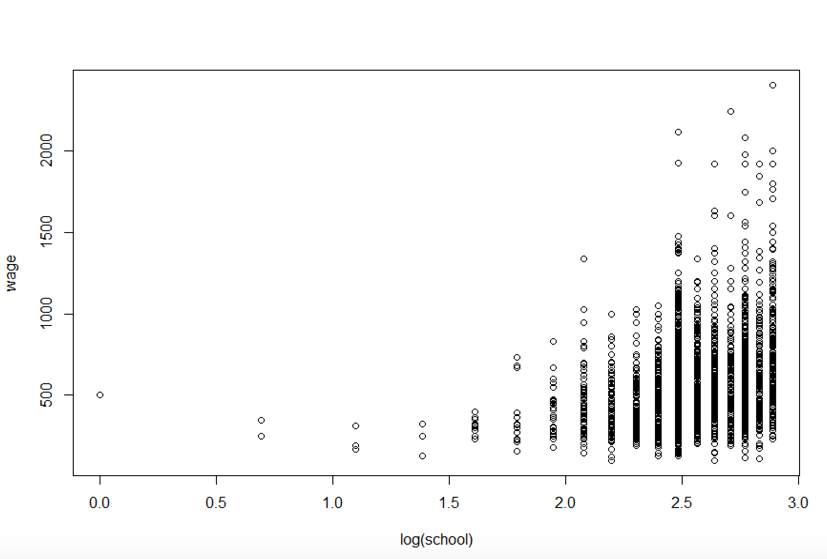


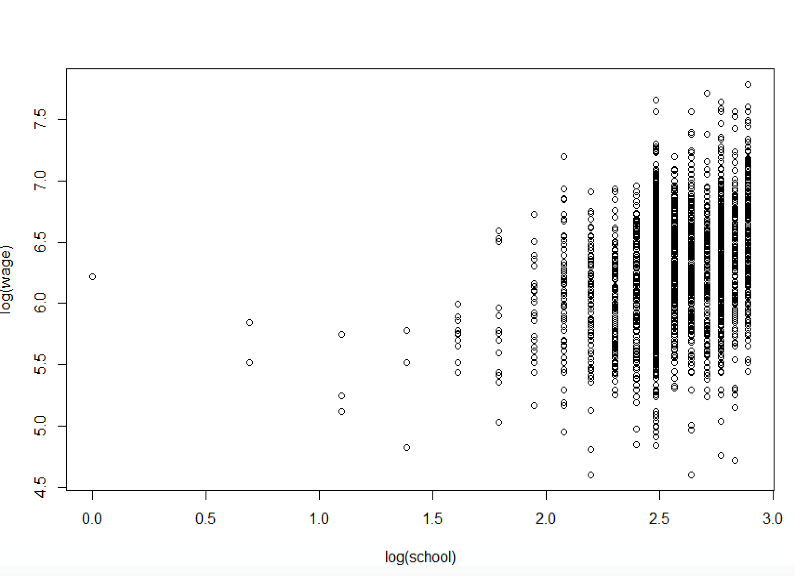
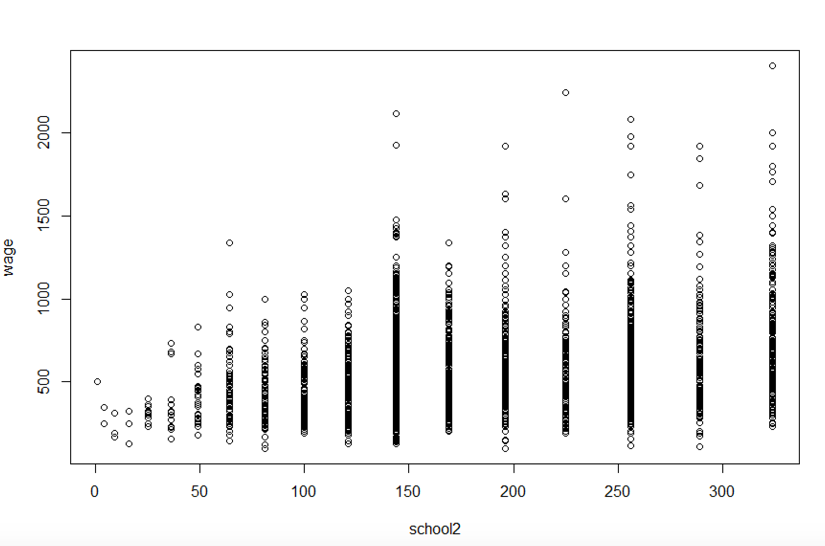


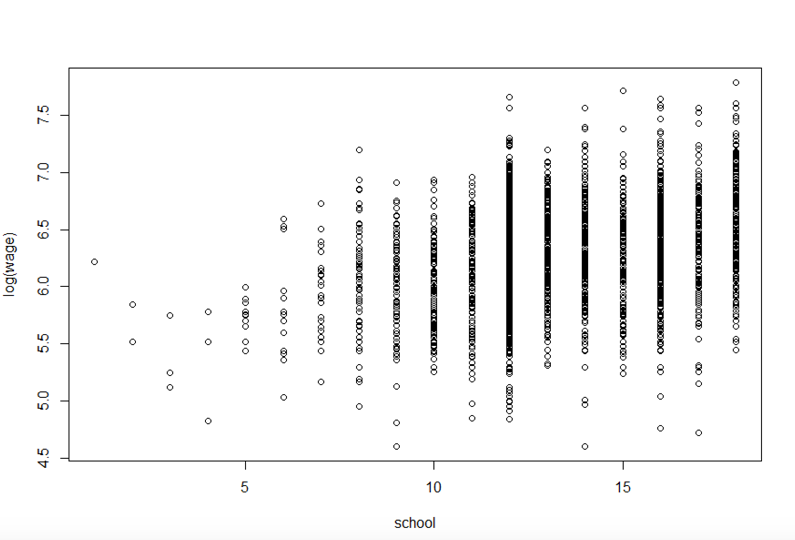
Conclusie: er is een specificatiefout.

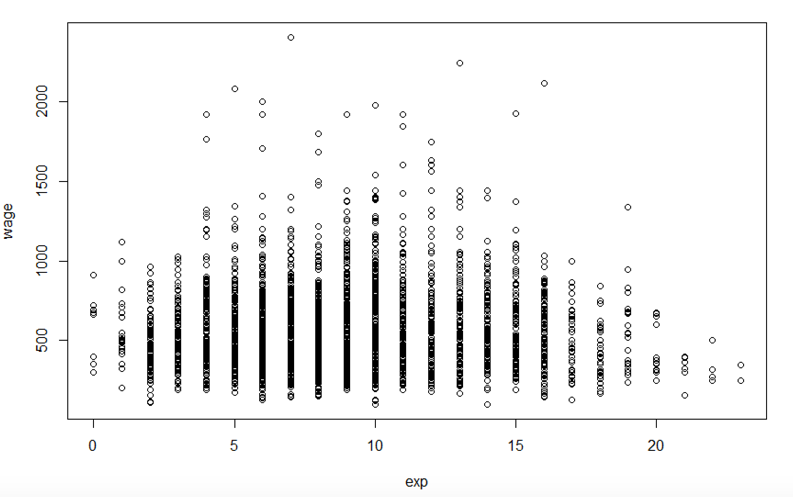
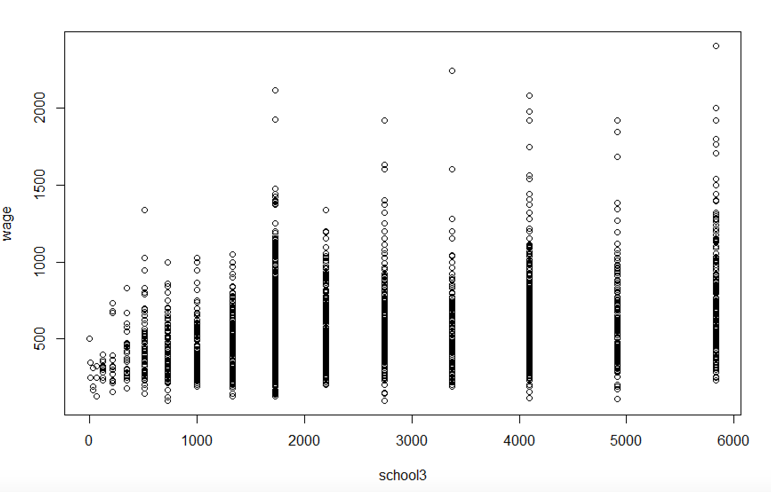
=> assumption NOT met

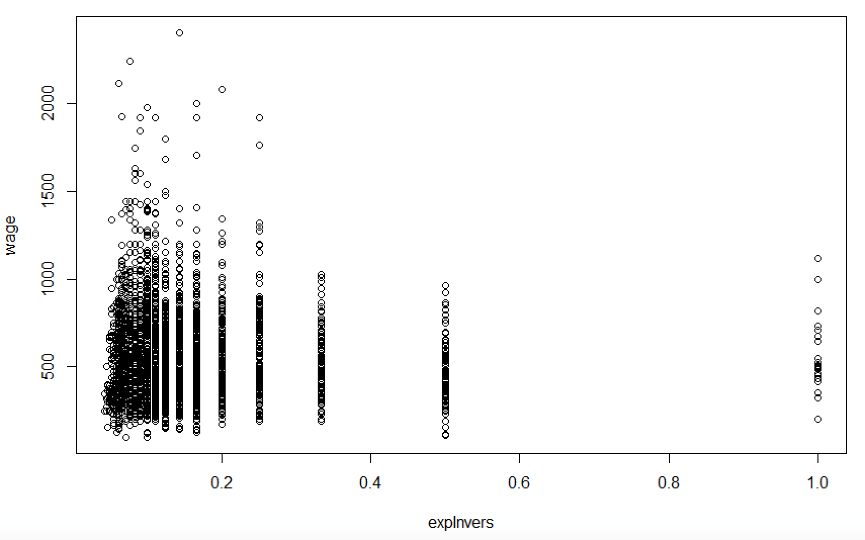
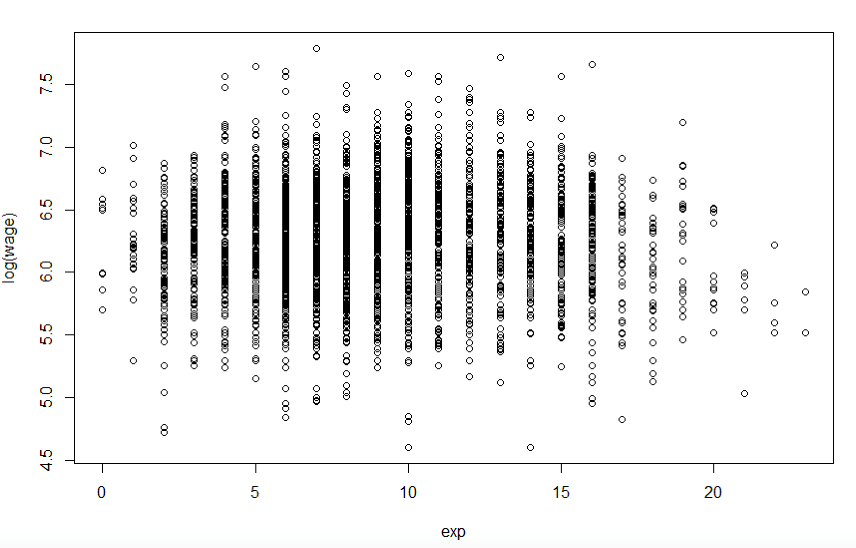
**Remedy**

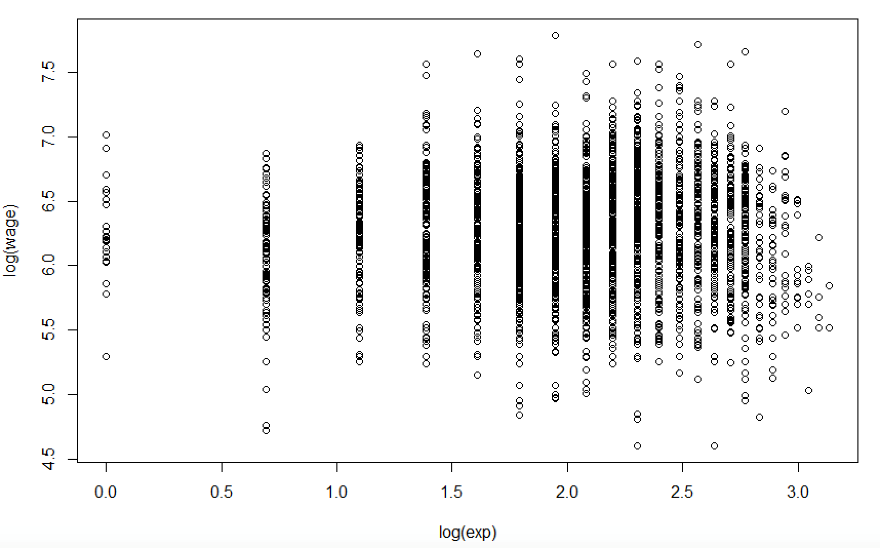


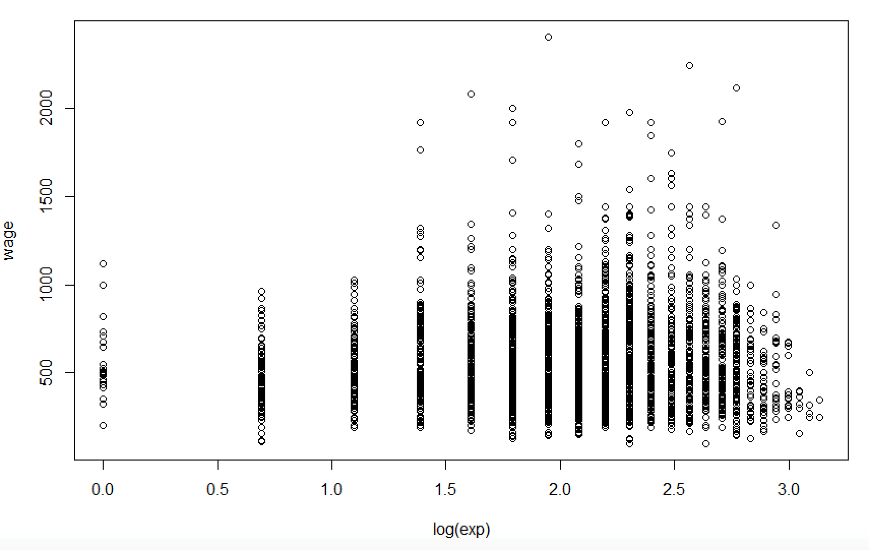


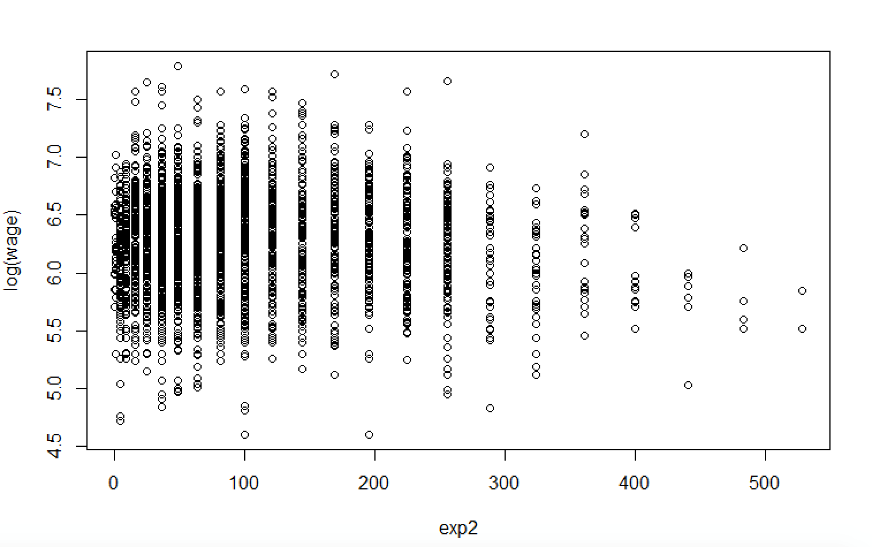




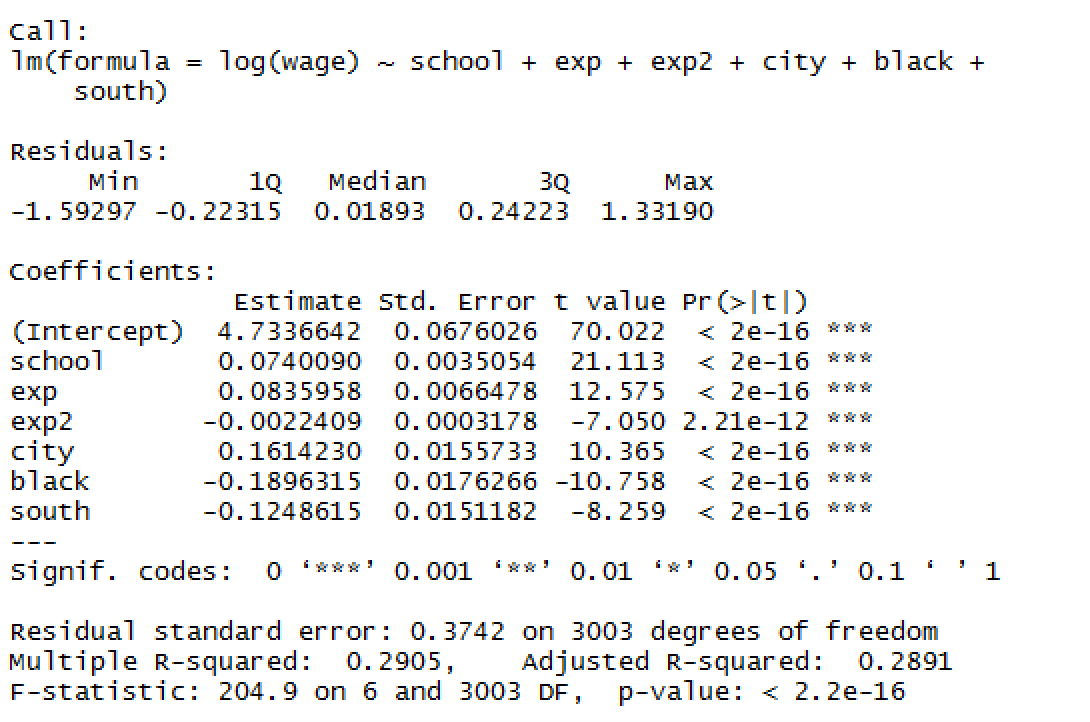




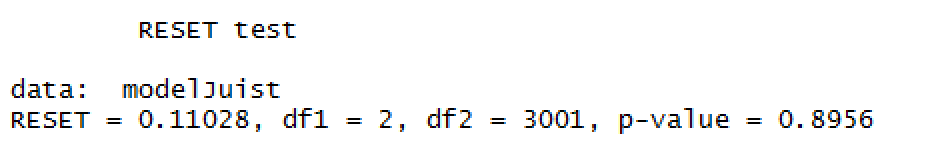




Dit lijkt ons het meest aangewezen model op basis van visuele inspectie van de plots.

****

We voerden reeds een Ramsey-Test uit die. De nulhypothese dat er geen specificatiefout is wordt bijgevolg niet verworpen.

****

Conclusie: er is geen specificatiefout.

## Assumption 10: Residuals are normally distributed

H0 : Residuals normaal verdeeld (Residuals ~N)

H1 : Residuals niet normaal verdeeld (Residuals≁N)

Jarque-Bera Normality Test

data: model$residuals

JB = 3687.7, p-value < 2.2e-16

alternative hypothesis: greater

p=0.0000 🡺 We verwerpen onze nulhypothese op 5% significantieniveau

**Afbeelding met schermafbeelding

Beschrijving is gegenereerd met zeer hoge betrouwbaarheid**

Conclusie: De storingstermen zijn niet normaal verdeeld.

=> assumption NOT met

# Step 4: Endogeneity + General remediation and conclusion

## Endogeneity

Model specification

https://lh5.googleusercontent.com/F2tVlvedqdpMF5BI24cEYbntH0vEvvz8wLG7PVqRN6BsY8y8pv_jU1owKp14NahL-w1DpLdNrBN7BvZh1_fHkSdnI93Qi4adPhY-GE41ASfbJEGcuTqnSPzoe8wIHmVOH2I83N6X

https://i.gyazo.com/827800730cbc9f9eedb6313a958149b7.png

https://i.gyazo.com/6df185f573262f05bf19e70410ec6986.png

Order condition

Schooling, wage en experience zijn de 3 endogene variabelen. Alle andere variabelen (nl. urbanization dummy, racial dummy, region dummy en near to college dummy) zijn predetermined. Uit de specificatie van het model kunnen we bijgevolg afleiden dat de instruments van de wage equation: near.to.college dummy en age zijn, daar deze niet in de vergelijking voorkomen.

M = 3, K = 5

Wage equation:

* m = 3, k = 3
* m + k = 6 = K + 1 = 6 => **exactly identified**
* instruments: near to college en age
* We mogen 2SLS toepassen

Schooling equation:

* m = 1, k = 4
* m + k = 5 < K + 1 = 6 => **over-identified**
* instruments: age
* We mogen 2SLS toepassen

Experience equation:

* m = 2, k = 1
* m + k = 3 < K + 1 = 6 => **over-identified**
* instruments: urbanization, near to college, racial & region
* We mogen 2SLS toepassen

Two Stage Least Squares

Call:

ivreg(formula = wage ~ school + city + black + south + exp |

distance + city + black + south + age)

Residuals:

Min 1Q Median 3Q Max

-879.1596 -187.1703 -0.7741 164.7838 1488.6584

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1231.124 349.406 -3.523 0.000432 \*\*\*

school 118.710 26.214 4.528 6.17e-06 \*\*\*

city 22.805 27.365 0.833 0.404704

black 11.721 40.193 0.292 0.770602

south -30.008 17.516 -1.713 0.086779 .

exp 25.634 1.745 14.687 < 2e-16 \*\*\*

Diagnostic tests:

df1 df2 statistic p-value

Weak instruments (school) 2 3004 10.71 2.31e-05 \*\*\*

Weak instruments (exp) 2 3004 2416.49 < 2e-16 \*\*\*

Wu-Hausman 1 3003 13.29 0.000271 \*\*\*

Sargan 0 NA NA NA

---

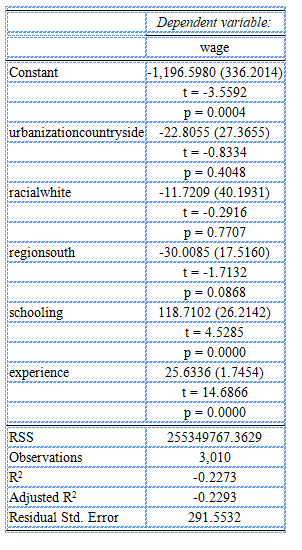
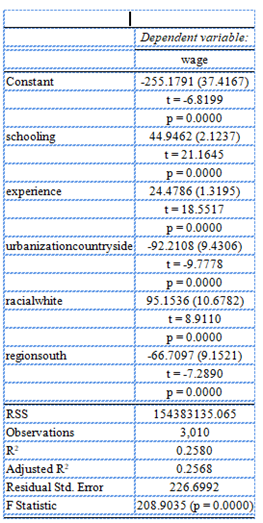
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 291.6 on 3004 degrees of freedom

Multiple R-Squared: -0.2273, Adjusted R-squared: -0.2293

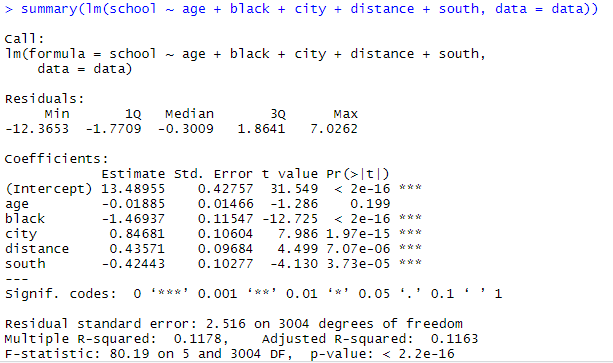
Wald test: 110 on 5 and 3004 DF, p-value: < 2.2e-16

We merken op via Wu-Hausman dat p<0.05 hierdoor kunnen we de nulhypothese dat er geen endogeniteit is verwerpen.

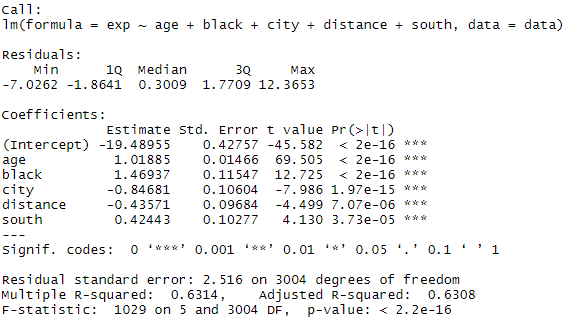


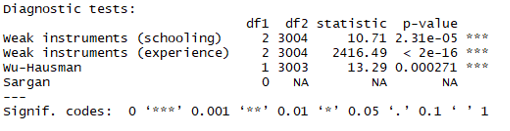
Wanneer we OLS met 2SLS vergelijken zien we zeer duidelijke verschillen wat in de richting van endogeniteit wijst.

Weak instruments



R² = 0.11 => wijst op weak instruments

  
R² = 0.63 => wijst op strong instruments maar is geen voldoende voorwaarde



Beide p-waardes zijn kleiner dan 0, wat wijst op weak instruments.

## General remediation and conclusion



Call:

lm(formula = log(wage) ~ school + exp + I(exp^2) + south + black +

city)

Residuals:

Min 1Q Median 3Q Max

-1.59297 -0.22315 0.01893 0.24223 1.33190

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.7336642 0.0676026 70.022 < 2e-16 \*\*\*

school 0.0740090 0.0035054 21.113 < 2e-16 \*\*\*

exp 0.0835958 0.0066478 12.575 < 2e-16 \*\*\*

I(exp^2) -0.0022409 0.0003178 -7.050 2.21e-12 \*\*\*

south -0.1248615 0.0151182 -8.259 < 2e-16 \*\*\*

black -0.1896315 0.0176266 -10.758 < 2e-16 \*\*\*

city 0.1614230 0.0155733 10.365 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3742 on 3003 degrees of freedom

Multiple R-squared: 0.2905, Adjusted R-squared: 0.2891

F-statistic: 204.9 on 6 and 3003 DF, p-value: < 2.2e-16

**Assumption 1: Linearity in the parameters**

The specification indicates linearity in the parameters. The assumption is fulfilled.

**Assumption 2: The X-values are determinstic**

The X-values are stochastic. This means that, when drawing a new sample, we get different values for X. The assumption is not fulfilled.

If the X’s are stochastic, then we must specify how the X’s and 𝜇𝑖 are distributed.

> cor.test(modelRem$residuals,school)

Pearson's product-moment correlation

data: modelRem$residuals and school

t = 1.661e-13, df = 3008, p-value = 1

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.035727 0.035727

sample estimates:

cor 3.028485e-15

> cor.test(modelRem$residuals,exp)

Pearson's product-moment correlation

data: modelRem$residuals and exp

t = 2.566e-14, df = 3008, p-value = 1

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.035727 0.035727

sample estimates:

cor 4.678674e-16

> cor.test(modelRem$residuals,exp^2)

Pearson's product-moment correlation

data: modelRem$residuals and exp^2

t = -6.7617e-16, df = 3008, p-value = 1

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.035727 0.035727

sample estimates:

cor -1.232872e-17

> cor.test(modelRem$residuals,south)

Pearson's product-moment correlation

data: modelRem$residuals and south

t = -9.6935e-16, df = 3008, p-value = 1

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.035727 0.035727

sample estimates:

cor -1.767433e-17

> cor.test(modelRem$residuals,city)

Pearson's product-moment correlation

data: modelRem$residuals and city

t = -1.2629e-15, df = 3008, p-value = 1

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.035727 0.035727

sample estimates:

cor -2.302708e-17

> cor.test(modelRem$residuals,black)

Pearson's product-moment correlation

data: modelRem$residuals and black

t = -6.0776e-15, df = 3008, p-value = 1

alternative hypothesis: true correlation is not equal to 0

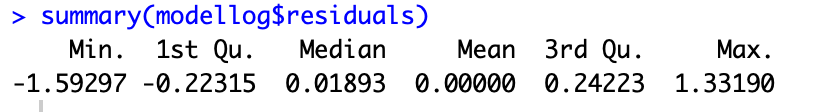
95 percent confidence interval:

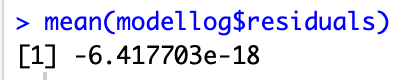
-0.035727 0.035727

sample estimates:

cor -1.108143e-16

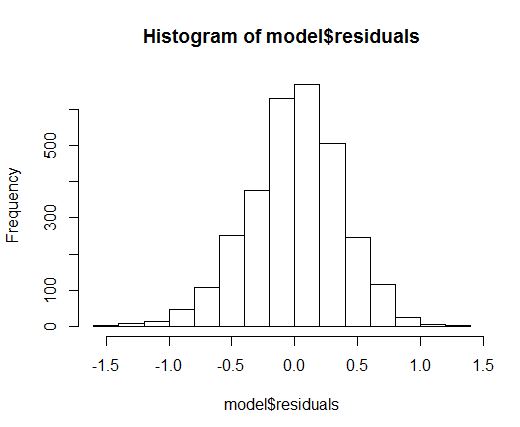
**Assumption 3: The (conditional) expectation of the error terms equals zero**

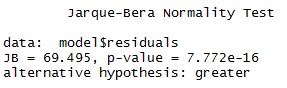




Uit deze gegevens kunnen we al afleiden dat het gemiddelde van de storingstermen 0 is met slechts kleine afwijkingen (minimum -1,59, max 1,33 = goed)

1 verdeling storingstermen





Uit de Jarque-Bera test kunnen we afleiden dat de storingstermen niet normaal verdeeld zijn. We verwerpen namelijk de nulhypothese op het 5% significantieniveau. Maar jarque bera is heel gevoelig voor afwijkingen als er weinig steekproeven zijn en/of te kleine grote, dus we kunnen wel nog steeds van een normale verdeling uitgaan adhv de figuur.



De kurtosis is groter dan 3, dus hebben we te maken met een gepiekte verdeling



De skewness ligt hier in het interval (-0.8, 0.8) hieruit kunnen we afleiden dat onze verdeling niet scheef is

**Assumption 4: The variances of the error terms is constant (homoscedasticity)**

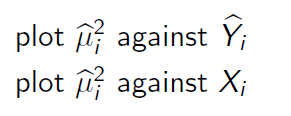
**Detection**

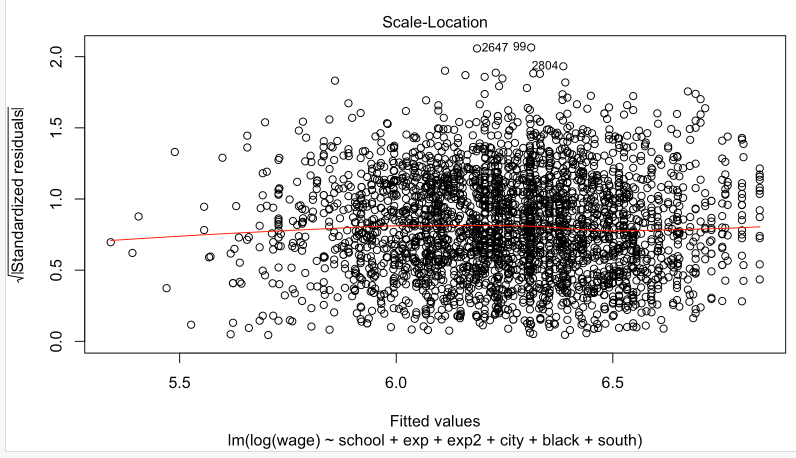
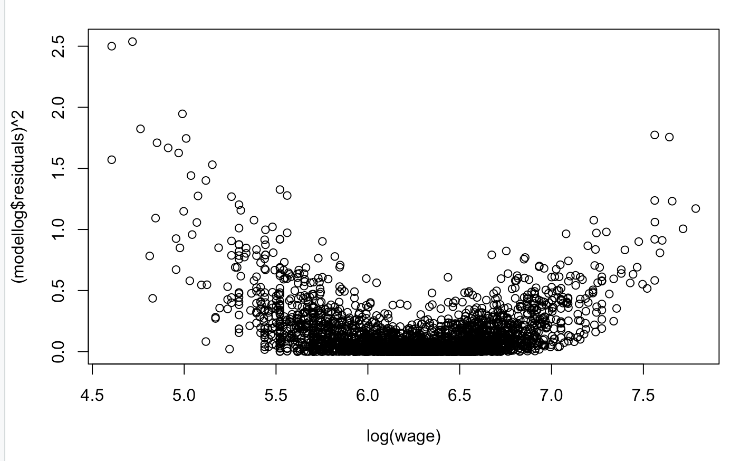
Informal

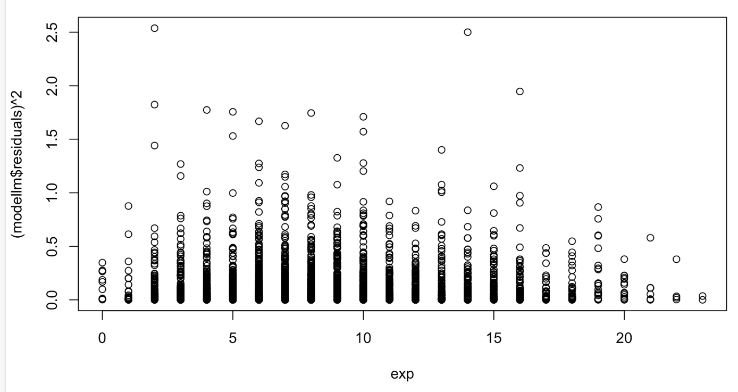
* Intuitive from the nature of the problem

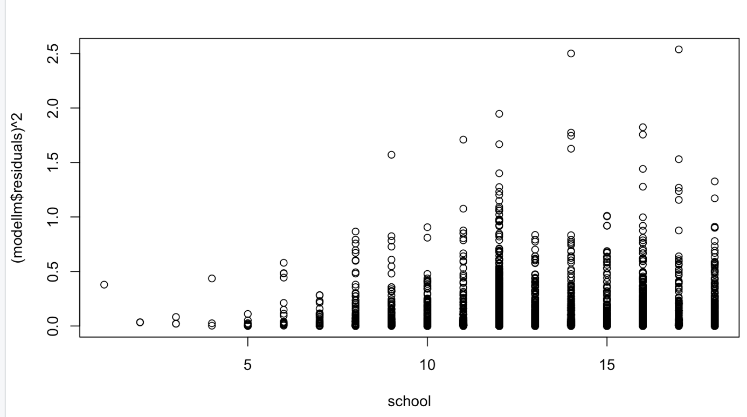
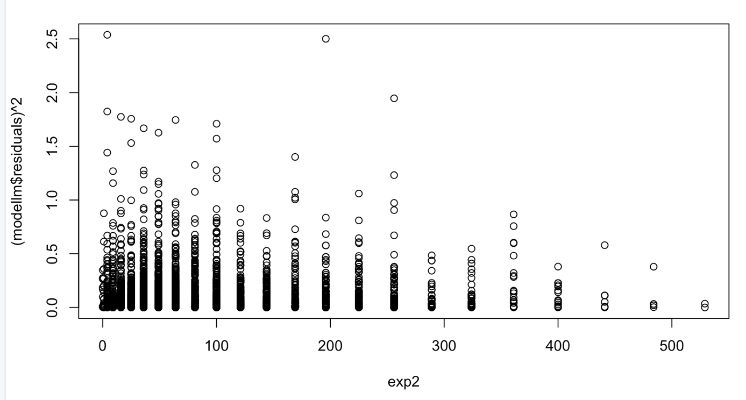
In cross-sections with heterogeneous entities, heteroskedasticity is probably the rule rather than the exception.

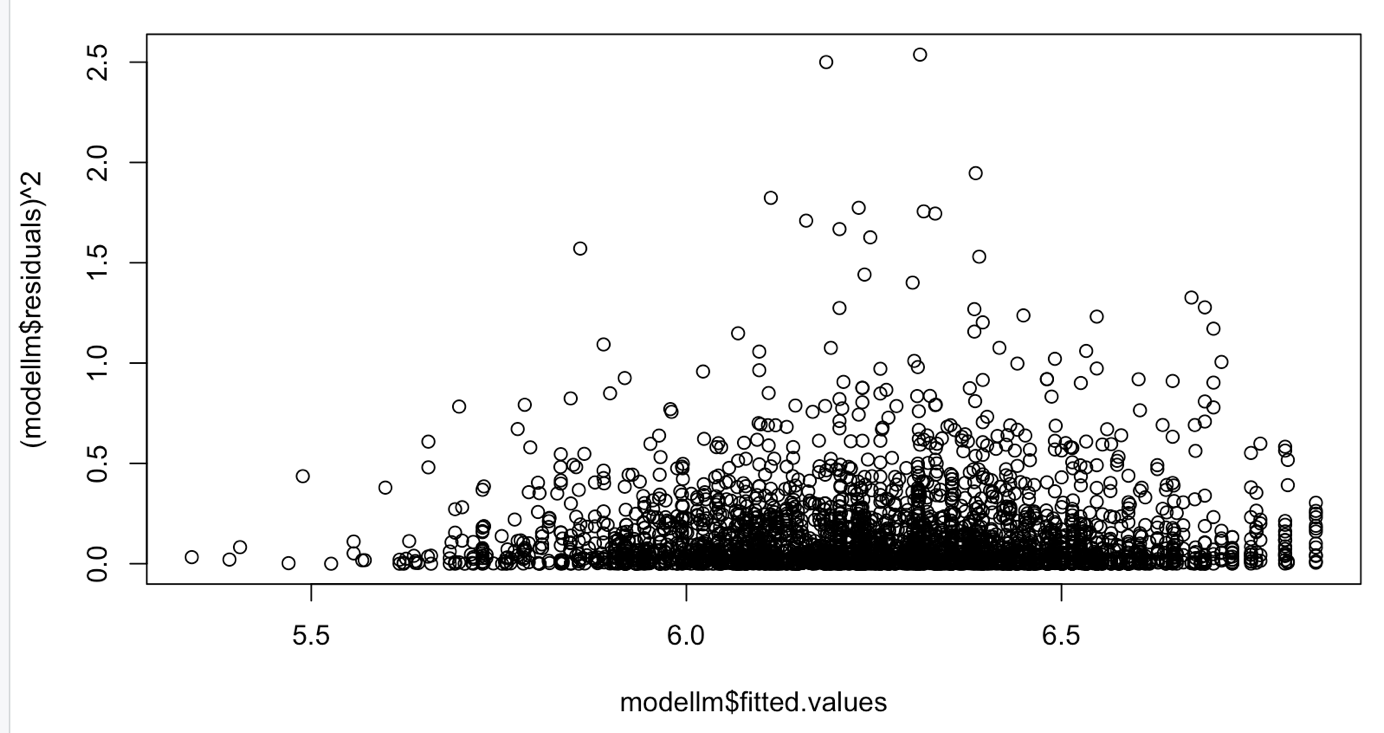
* 



Afbeelding met schermafbeelding, kaart

Automatisch gegenereerde beschrijving





Formal

1. **Goldfeld-Quandt test**

Assumption 1: The error terms are normally distributed

The error terms are in this case not normally distributed, but asymptotic theory: OLS estimator tends to the normal distribution when n -> .

Assumption 2: is positively related to one of the explanatory variables

If we look at the plots above, is positively related to schooling, but the relation with experience is more difficult.

> gqtest(modelRem, fraction = 10,alternative = "two.sided", order.by=exp)

Goldfeld-Quandt test

data: modelRem

GQ = 0.93036, df1 = 1493, df2 = 1493, p-value = 0.1632

alternative hypothesis: variance changes from segment 1 to 2

> gqtest(modelRem, fraction = 10,alternative = "greater", order.by=exp)

Goldfeld-Quandt test

data: modelRem

GQ = 0.93036, df1 = 1493, df2 = 1493, p-value = 0.9184

alternative hypothesis: variance increases from segment 1 to 2

> gqtest(modelRem, fraction = 10,alternative = "less", order.by=exp)

Goldfeld-Quandt test

data: modelRem

GQ = 0.93036, df1 = 1493, df2 = 1493, p-value = 0.08162

alternative hypothesis: variance decreases from segment 1 to 2

> gqtest(modelRem, fraction = 10,alternative = "two.sided", order.by=school)

Goldfeld-Quandt test

data: modelRem

GQ = 1.0404, df1 = 1493, df2 = 1493, p-value = 0.4448

alternative hypothesis: variance changes from segment 1 to 2

> gqtest(modelRem, fraction = 10,alternative = "greater", order.by=school)

Goldfeld-Quandt test

data: modelRem

GQ = 1.0404, df1 = 1493, df2 = 1493, p-value = 0.2224

alternative hypothesis: variance increases from segment 1 to 2

> gqtest(modelRem, fraction = 10,alternative = "less", order.by=school)

Goldfeld-Quandt test

data: modelRem

GQ = 1.0404, df1 = 1493, df2 = 1493, p-value = 0.7776

alternative hypothesis: variance decreases from segment 1 to 2

1. **White’s General Heteroskedasticity test**

> expXschool <- exp\*school

> expXblack <- exp\*black

> expXsouth <- exp\*south

> expXcity <- exp\*city

> expXexp2 <- exp\*exp\*exp

> schoolXblack <- school\*black

> schoolXsouth <- school\*south

> schoolXcity <- school\*city

> schoolXexp2<-school\*exp\*exp

> blackXsouth <- black\*south

> blackXcity <- black\*city

> blackXexp2<-black\*exp\*exp

> southXcity <- south\*city

> southXexp2<-south\*exp\*exp

> cityXexp2<-city\*exp\*exp

> bptest(modelRem, ~ school+I(school^2)+exp+I(exp^2)+I(exp^2)+I(exp^4)+black+I(black^2)+south+I(south^2)+city+I(city^2)+expXschool+expXblack+expXsouth+expXcity+schoolXblack+schoolXsouth+schoolXcity+blackXsouth+blackXcity+southXcity+expXexp2+schoolXexp2+blackXexp2+southXexp2+cityXexp2)

studentized Breusch-Pagan test

data: modelRem

BP = 33.398, df = 23, p-value = 0.07438

> bptest(modelRem, ~ school+I(school^2)+exp+I(exp^2)+I(exp^2)+I(exp^4)+black+I(black^2)+south+I(south^2)+city+I(city^2))

studentized Breusch-Pagan test

data: modelRem

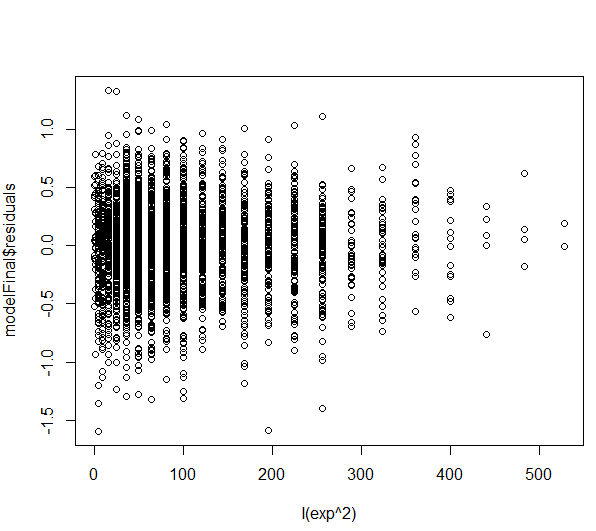
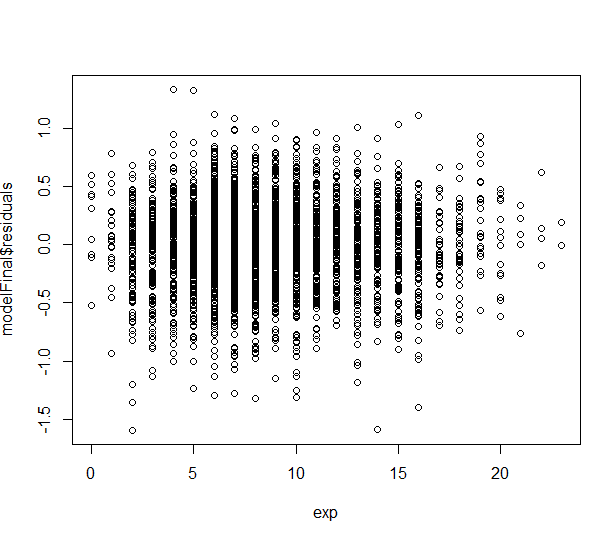
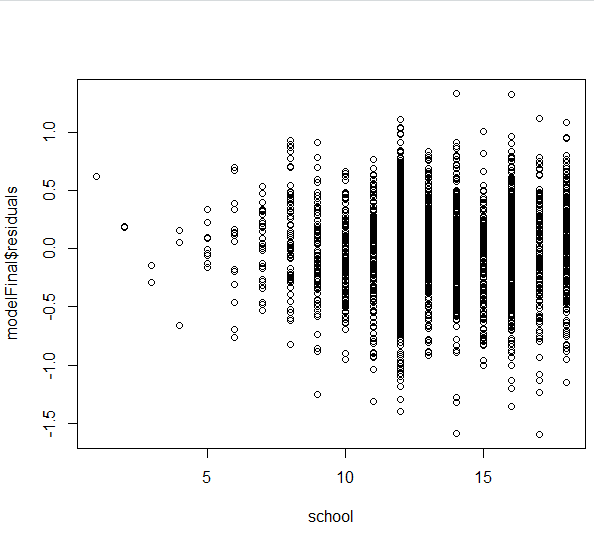
BP = 7.1598, df = 8, p-value = 0.5195

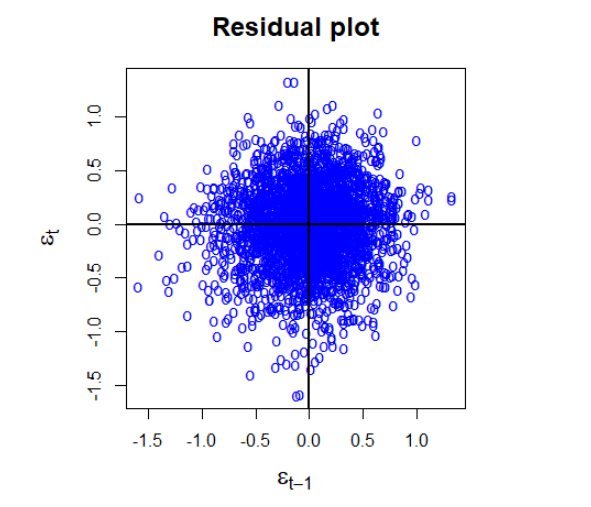
Don’t reject H0 (dus wel homoscedastisch)

**Assumption 5: There is no correlation in the error terms (no autocorrelation)**

**Detection:**

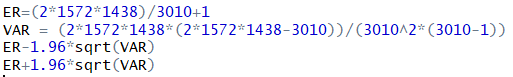
1. **Graphical method**
2. We plot the estimated error terms against schooling, experience and exp²:

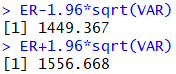


1. Plot the estimated error terms against the estimated error terms of the previous observation 

Deze plot wijst niet op autocorrelatie

1. **Tests**
2. **Runs test**





1448 (= observed runs) lies outside the interval [1449.367-1556.668]: Although it’s very close, we can reject H0 and conclude that there might be autocorrelation in our model.

1. **Durbin-Watson d test**

Assumption 1: X’s are deterministic.

In our case the X’s are stochastic.

Assumption 2: AR(1) pattern in the error terms. The d statistic cannot be used to detect higher-order autoregressive schemes.

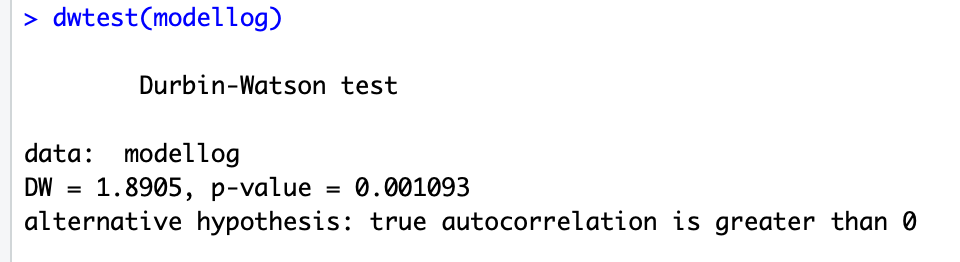
Assumption3: Error terms normally distributed

The error terms are in this case not normally distributed, but asymptotic theory: OLS estimator tends to the normal distribution when n -> .

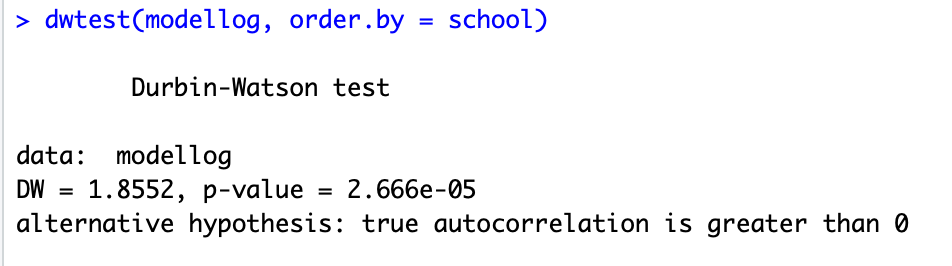
Assumption 4: The regression model does not include the lagged values of the dependent variable as one of the explanatory variables.

Autocorrelation relevant for time series and panel data (logical ordering in the data). We will order the data according to experience and schooling

Ordered by exp



Ordered by schooling

****

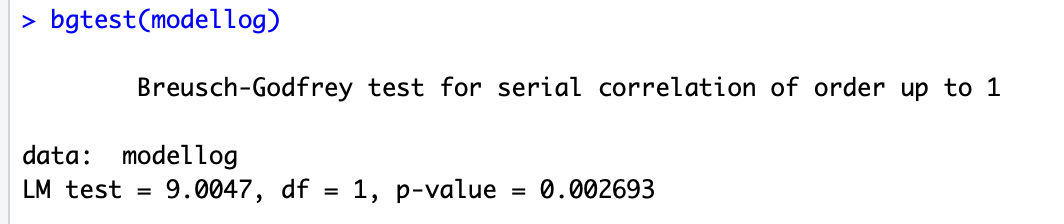
1. **Breusch-Godfrey LM test:**

Advantages:

* X’s are allowed to be stochastic
* Regression model is allowed to include lagged dependent variables
* Error terms not restricted to be generated by AR(1) process

Disadvantages:

* We need to choose p (We will test for first-order (p=1) and second-order (p=2) serial correlation)
* Asymptotic test

****

**Afbeelding met schermafbeelding

Automatisch gegenereerde beschrijving**

1. Cochrane Orcutt iteration (EGLS)**Afbeelding met schermafbeelding

   Automatisch gegenereerde beschrijving**

Mss beter output van slimme groepje nemen?

**REMEDIAL MEASURES: take autocorrelation into account**

1. Correction variance OLS (Newey-West correction)

Correction variance OLS (Newey-West correction)

Heteroskedasticity and autocorrelation consistent standard errors: If a sample is reasonably large, one should use the Newey-west procedure to correct OLS standard errors not only in situations of autocorrelation but also in cases of heteroscedasticity, for the HAC method can handle both.

Assume n -> ∞; The Newey-west procedure is strictly speaking valid in large samples and may not be appropriate in small samples.

Does not assume a particular pattern

Afbeelding met tekst

Automatisch gegenereerde beschrijving

**Assumptie 6:**

aantal observaties > aantal te schatten paramers: ok

**Assumptie 7:**

variantie in de x-waarden: ok

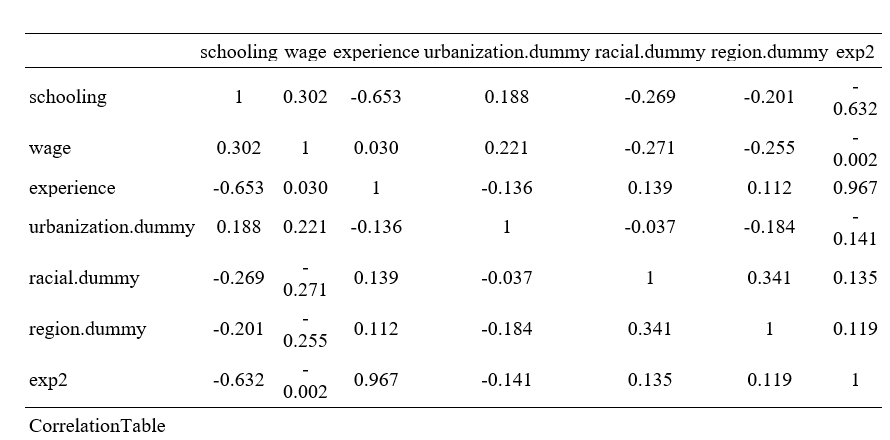
**Assumptie 8:**

multicollineariteit

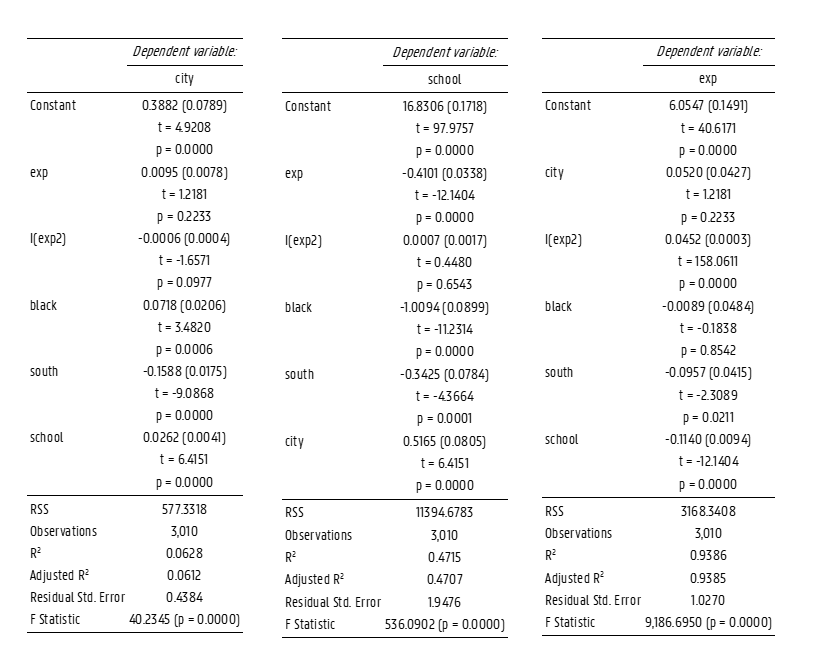
1ste methode kijk naar R^2 en t waarden van de variabelen (zie default hier net boven)

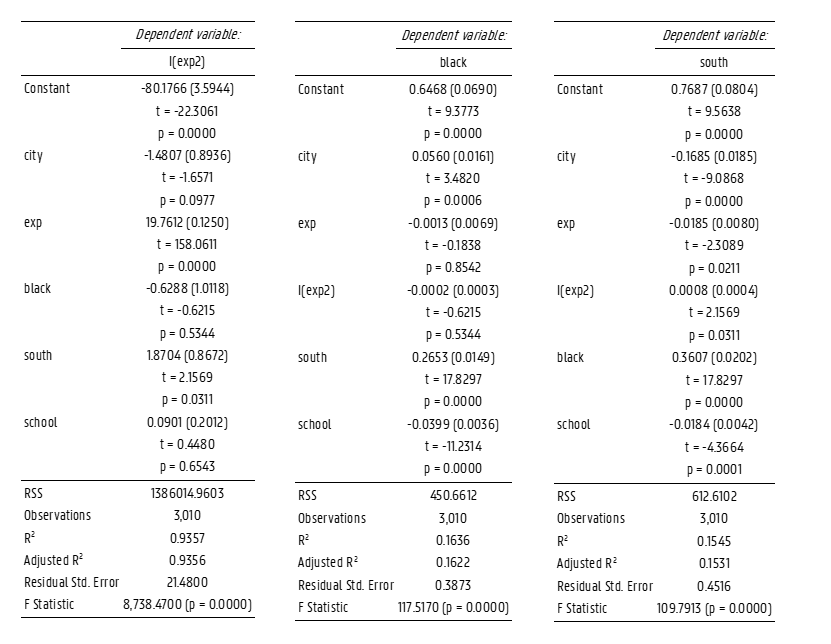
R^2 model = 0.2905 en al de t-waarden zijn voldoende groot (zeer significant) 🡪 hier geen reden om multicollineariteit te verwachten

2de  methode: pairwise correlation between explanatory variables (sufficient but not necessary)



3de methode: estimate auxillary regressions





VIF berekenen adhv R^2 in de auxillary regressions

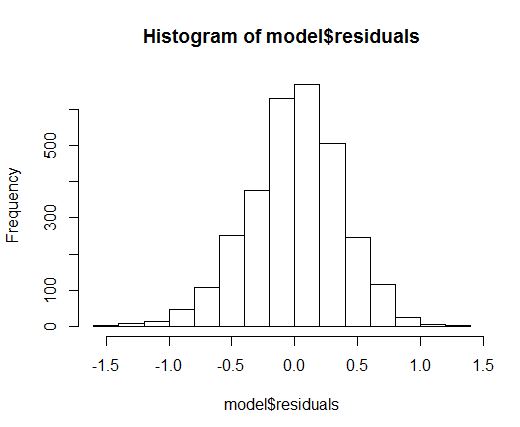
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variabele | school | city | black | south | exp | exp2 |
| R^2 | 0.4715 | 0.06277 | 0.1636 | 0.1545 | 0.9386 | 0.9357 |
| VIF | 1.89 | 1.066 | 1.20 | 1.18 | 16.29 | 15.55 |

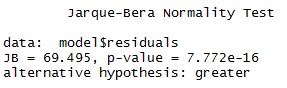


Alle VIF’s buiten buiten die voor exp en exp^2 zijn in orde… dit komt omdat exp en exp^2 natuurlijk veel van elkaars variantie verklaren als we R^2 berekenen zonder deze 2 verklarende variabelen in elkaars regressie dan bekomen we een R^2 van 0,4281 voor exp 🡪 VIF = 1,75 en R^2 van 0,4007 voor exp ^2 🡪 VIF = 1,67 dus deze zijn dan wel onder de grens van 4 dus geen gevaar voor multicollineariteit

**Assumption 9: The error terms are normally distributed**

1 verdeling storingstermen (adhv histogram)



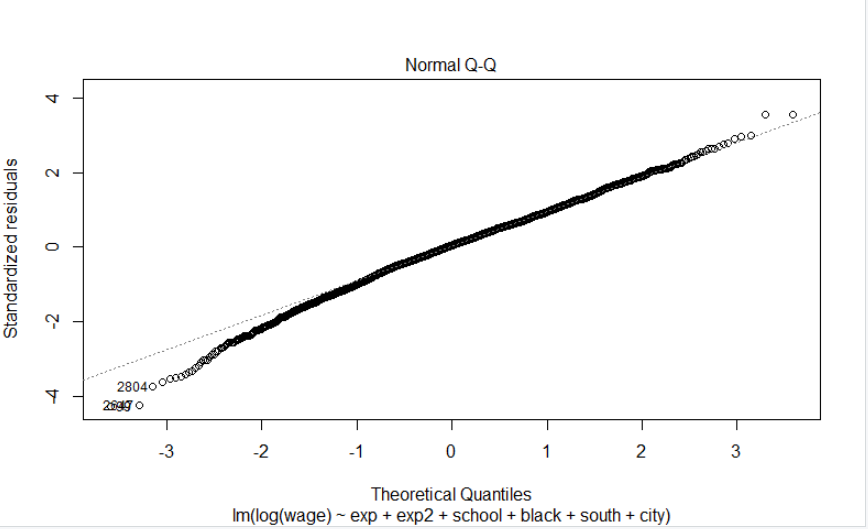


Uit de Jarque-Bera test kunnen we afleiden dat de storingstermen niet normaal verdeeld zijn. We verwerpen namelijk de nulhypothese op het 5% significantieniveau. Maar jarque bera is heel gevoelig voor afwijkingen als er weinig steekproeven zijn en/of te kleine grote, dus we kunnen wel nog steeds van een normale verdeling uitgaan adhv de figuur.



De kurtosis is groter dan 3, dus hebben we te maken met een gepiekte verdeling



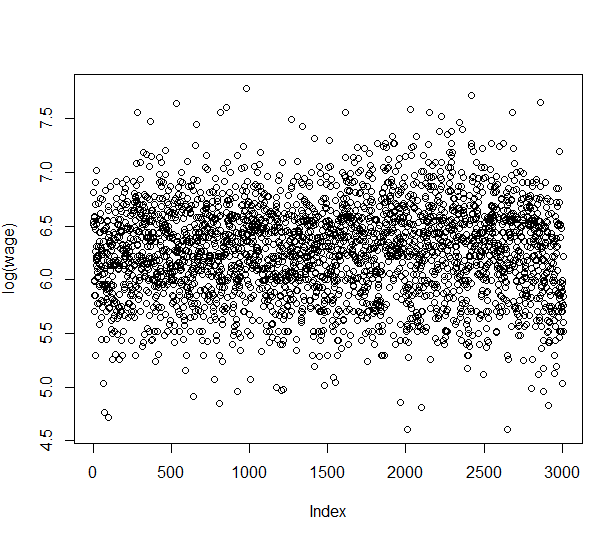


De skewness ligt hier in het interval (-0.8, 0.8) hieruit kunnen we afleiden dat onze verdeling niet scheef is (maar het is negatief dus kans op afwijking aan linkerstaart) . adhv QQ plot kunnen we linksscheve verdeling vaststellen.

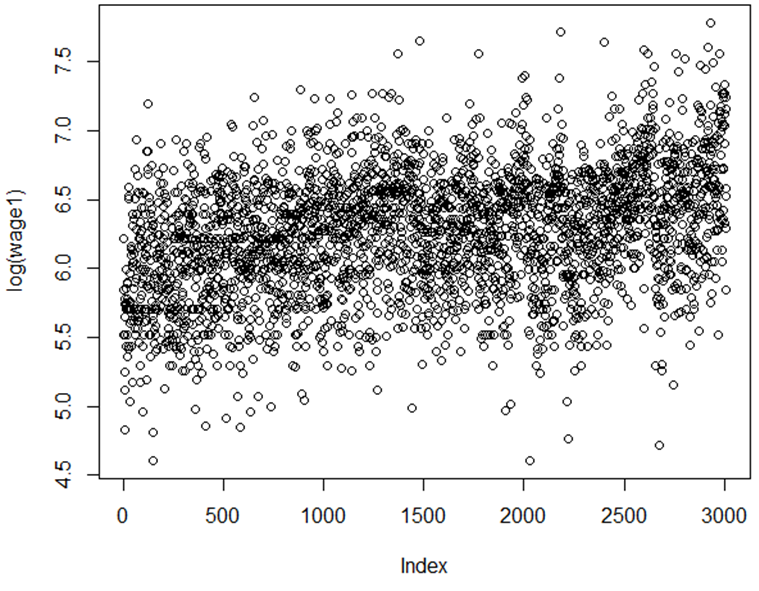
**Assumption 10: The model is correctly specified, so there is no specification bias.**

**Detection:**

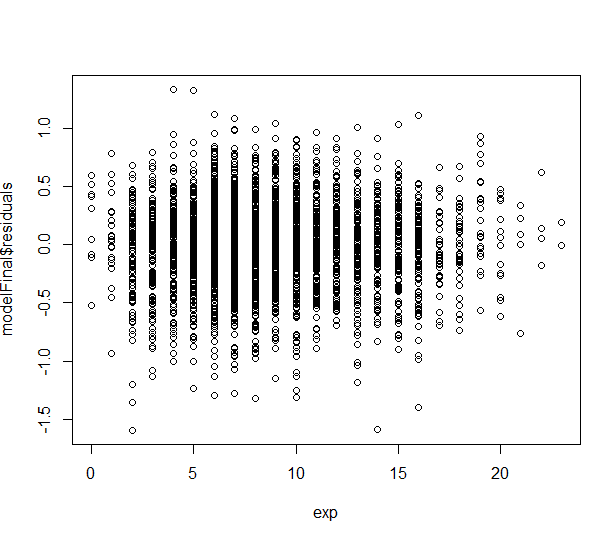
* Actual values ordered by the explanatory variable experience



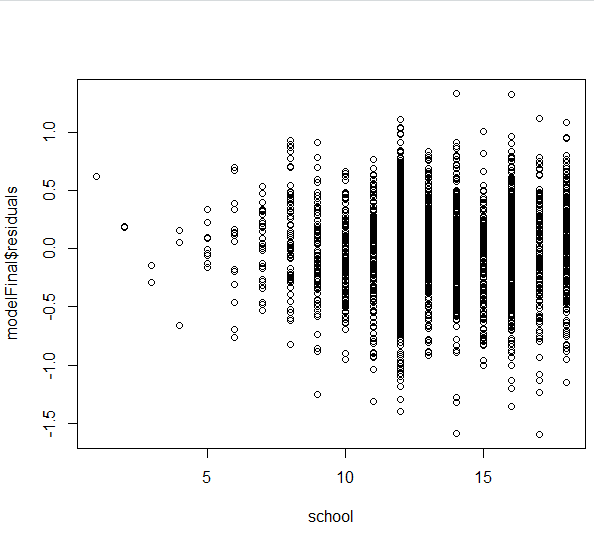
* Actual values ordered by the explanatory variable schooling

****

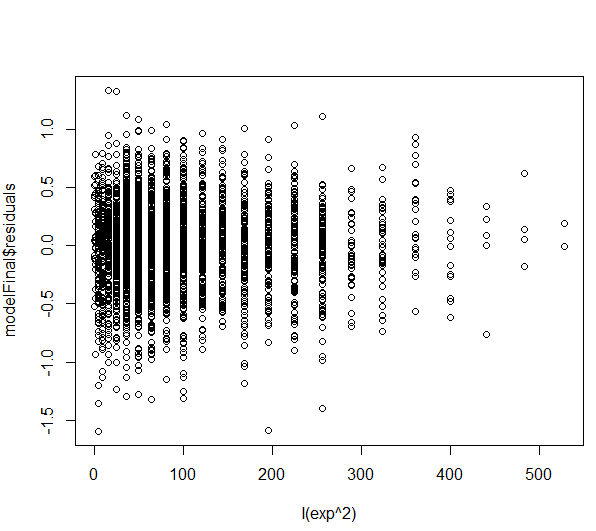
* **Visual inspection of the error terms**
* Order the error terms (based on time or explanatory variable)
* Systematic pattern? If there are specification errors, the residuals will exhibit noticeable patterns
* Error terms ordered by explanatory variable experience



* Error terms ordered by explanatory variable schooling



* Error terms ordered by explanatory variable experience²



**TESTS**

1. Overfit the model and use standard t tests and F tests to determine significance

Start met deze specificatie: ****

Bekijk coeff vd regressive:

**Afbeelding met tekst

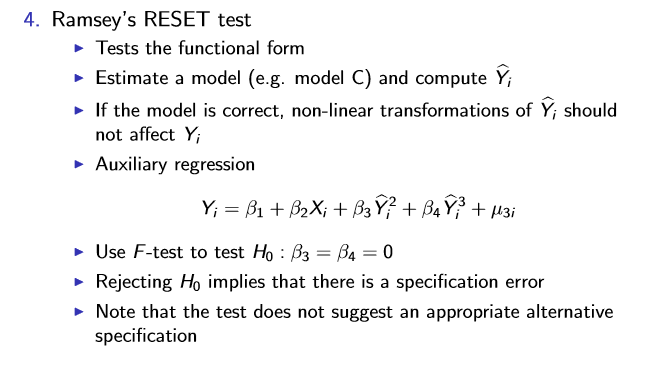
Automatisch gegenereerde beschrijving**

haal ‘age’ er uit want data is NA in deze regressie

Afbeelding met schermafbeelding

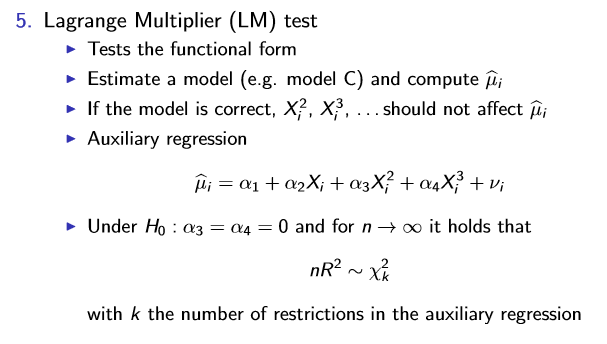
Automatisch gegenereerde beschrijving

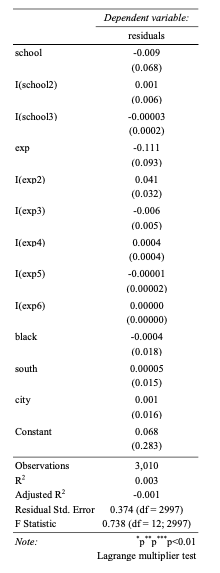
Alle coefficienten op near.to.college zijn zeer significant (p<0.0001) buiten near to college p tussen 1 en 5%, dus wel significant op 5% maar voor een goede modellering willen we liefst een signifiacntie tot op 1% of lager, daarom halen we near.to.college er ook uit.

Afbeelding met schermafbeelding

Automatisch gegenereerde beschrijving

1. Lagrange Multiplier (LM) test





R² is 0,003. Now n\*R² = 3010\*0.003=8.87 (LM in R code) which has, asymptotically, a chi-square distribution with 6 df. The 5 percent critical chi-square value for 6 df is 12.5916. We do not reject H0. We conclude that there is no specification error in our model.

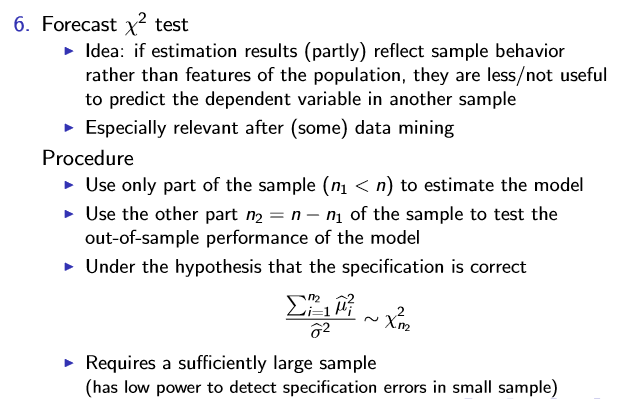


R² is 0.003. Now n\*R² = 3010\*0.002=8.83 (LM in R code) which has, asymptotically, a chi-square distribution with 4 df. The 5 percent critical chi-square value for 4 df is 9.48773. We do not reject H0. We conclude that there is no specification error in our model.



R² is 0.0002. Now n\*R² = 3010\*0.0002=0.54(LM in R code) which has, asymptotically, a chi-square distribution with 2 df. The 5 percent critical chi-square value for 2 df is 5,99 . We do not reject H0. We conclude that there is no specification error in our model.

1. Forecast χ² test



**Endogeneity**

1. Order condition:

M = 3 (Wi, Si, Ei)

K = 5 (Ui, Rai, Rei, Zi1, Age i)

* **Wage equation:**

m=3 (Wi, Si, Ei) , k=3 (Ui, Rai, Rei)

m+k=6 =K+1=6 🡪 exactly identified

* **Schooling equation:**

m=1 (Si) , k=4 (Ui, Rai, Rei, Zi1)

m+k=5<K+1=6 🡪 over-identified

* **Experience equation:**

m=2, k=1

m+k=3<K+1=6 🡪 over-identified

* **Exp² equation**

m=2, k=1

m+k=3<K+1=6 🡪 over-identified

**2SLS**

1. Bereken de proxy’s voor alle endogene variabelen door de endogene variabele te regressen op alle predetermined variables
2. **School**

**Afbeelding met tekst

Automatisch gegenereerde beschrijving**

1. **Exp**

**Afbeelding met tekst

Automatisch gegenereerde beschrijving**

1. **Exp2**

**Afbeelding met tekst

Automatisch gegenereerde beschrijving**

1. **2SLS met ivreg**

Afbeelding met tekst

Automatisch gegenereerde beschrijving

> twoSls2 <- ivreg(formula=log(wage)~school+exp+I(exp^2)+south+black+city|I(exp^2)+city+black+south+distance+age)

> summary(twoSls2,diagnostics = T)

Call:

ivreg(formula = log(wage) ~ school + exp + I(exp^2) + south +

black + city | I(exp^2) + city + black + south + distance +

age)

Residuals:

Min 1Q Median 3Q Max

-15.6889 -1.2847 0.7521 1.9962 5.5506

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.40314 11.04942 0.761 0.447

school 0.51175 1.31698 0.389 0.698

exp -2.28095 7.11258 -0.321 0.748

I(exp^2) 0.11777 0.36100 0.326 0.744

south -0.12615 0.11456 -1.101 0.271

black 0.53574 2.18594 0.245 0.806

city -0.08293 0.74440 -0.111 0.911

Diagnostic tests:

df1 df2 statistic p-value

Weak instruments (school) 2 3003 4446.825 <2e-16 \*\*\*

Weak instruments (exp) 2 3003 152.458 <2e-16 \*\*\*

Wu-Hausman 1 3002 6.351 0.0118 \*

Sargan 0 NA NA NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.834 on 3003 degrees of freedom

Multiple R-Squared: -39.69, Adjusted R-squared: -39.77

Wald test: 3.586 on 6 and 3003 DF, p-value: 0.001522