Økonometrieksamen

Sofie Teisen, Rebekka Hansen, Josefine Østergaard, Jonas Elkjær

13. June 2024

Contents

Eksamenssæt 1 **4**

Opgave 2 - Udfør grafisk modelkontrol 4

Opgave 3 - Test for heteroskedasticitet vha. Breusch-Pagan-testen og specialudgaven af White-testen 6

Opgave 4 - Beregn robuste standardfejl for modellen og sammenlign med resultaterne i spørgsmål 1 8

Opgave 5 - Test hypotesen H0: $\beta_2 = 1$ mod alternativet H1: $\beta_2 \neq 1$ 9

Opgave 6 - Test hypotesen H0: $\beta_3 = \beta_4 = 0$ 9

Opgave 7 - Estimer modellen vha. FGLS og kommenter på resultaterne 10

Opgave 8 - Har FGLS estimationen taget højde for al heteroskedasticiteten? 12

Eksamenssæt 2 **14**

Opgave 1 - Estimer de to modeller vha. OLS. Kommenter på outputtet, sammenlign og fortolk resultaterne. 14

Opgave 2 - Udfør grafisk modelkontrol af de to modeller. Hvilken model vil du foretrække? 15

Opgave 3 - Undersøg om de to modeller er misspecificerede vha. RESET-testet 17

Opgave 4 - Forklar hvorfor det kunne være relevant at medtage educ2 som forklarende variabel i de to modeller. Estimer de to modeller igen hvor educ2 inkluderes (med tilhørende koefficient $^{-}5$), kommenter kort på outputtet og udfør RESET-testet igen. 19

Opgave 5 - Test hypotesen $H0 : \beta_1 = \beta_5 = 0$ i begge modeller (fra spørgsmål 4). 21

Opgave 6 - Kunne der være problemer med målefejl i de to modeller? I hvilke tilfælde vil det udgøre et problem? 22

Eksamenssæt 3: Instrumentvariable **23**

Opgave 1 - Estimer modellen vha. OLS og kommenter på resultaterne 23

Opgave 2 - Hvorfor kunne vi være bekymrede for at uddannelse er endogen? 23

Opgave 3 - Er siblings, meduc og feduc brugbare som instrumenter? 24

Opgave 4 - Test om uddannelse er endogen 24

Opgave 5 - Estimer modellen vha. 2SLS hvor du gør brug af de tre beskrevne instrumenter. Sammenlign med resultaterne i spørgsmål 1. 25

Opgave 6 - Udfør overidentifikationstestet. Hvad konkluderer du? 26

Opgave 7 - Udfør hele analysen igen hvor du kun bruger meduc og feduc som instrumenter. Ændrer det på dine konklusioner? 26

Eksamenssæt 4 **29**

Opgave 1 - Opstil en lineær regressionsmodel for *participation* hvor du bruger de beskrevne forklarende variable. 29

(a) - Estimer modellen vha. OLS og kommenter på resultaterne. 29

(b) - Test om den partielle effekt af uddannelse er forskellig fra nul. 30

(c) - Test om den partielle effekt af alder er forskellig fra nul. 30

Opgave 2 - Opstil både en logit- og en probit-model for par t i cipat ion hvor du bruger de beskrevne forklarende variable. 31

(a) - Estimer modellerne. 31

(b) - Test om den partielle effekt af uddannelse er forskellig fra nul. 32

(c) - Test om den partielle effekt af alder er forskellig fra nul vha. et likelihoodratio-test. . . . 32

Opgave 3 - Vi vil gerne sammenligne den partielle effekt af *income* på tværs af modellerne. Beregn average partial effect (APE) og kommenter på resultaterne. 32

Opgave 4 - Vi vil gerne sammenligne den partielle effekt af *foreign* på tværs af modellerne. Beregn APE og kommenter på resultaterne. 33

Opgave 5 - Hvorfor er APE at foretrække frem for partial effect at the average (PEA)? 33

Opgave 6 - Sammenlign modellernes evne til at prædiktere ved at beregne percent correctly predicted for hver model. 33

Eksamenssæt 1

MANGLER MINIMISERING AF SSR OG FORKLARING AF DENNE ## Opgave 1 - Estimer modellen vha. OLS. Kommenter på outputtet og fortolk resultaterne For at estimere modellen vha. OLS, opstilles der en regression. For at estimere modellen antages det at $E(u) = 0$ og $Cov(u, x) = 0$ hertil

```
model1 = lm(logsal ~ educ+logbegin+male+minority)
summary(model1)
```

```
##
## Call:
## lm(formula = logsal ~ educ + logbegin + male + minority)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.455724 -0.115077 -0.005164  0.107650  0.870602
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  0.8486767  0.0751152 11.2983 < 0.00000000000000022 ***
## educ         0.0232683  0.0038696  6.0131        0.000000003664 ***
## logbegin     0.8217988  0.0360314 22.8078 < 0.00000000000000022 ***
## male         0.0481556  0.0199103  2.4186               0.01596 *
## minority    -0.0423686  0.0203417 -2.0828               0.03781 *
## ---
## 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:  0.80412,    Adjusted R-squared:  0.80245
## F-statistic: 481.32 on 4 and 469 DF,  p-value: < 0.000000000000000222
```

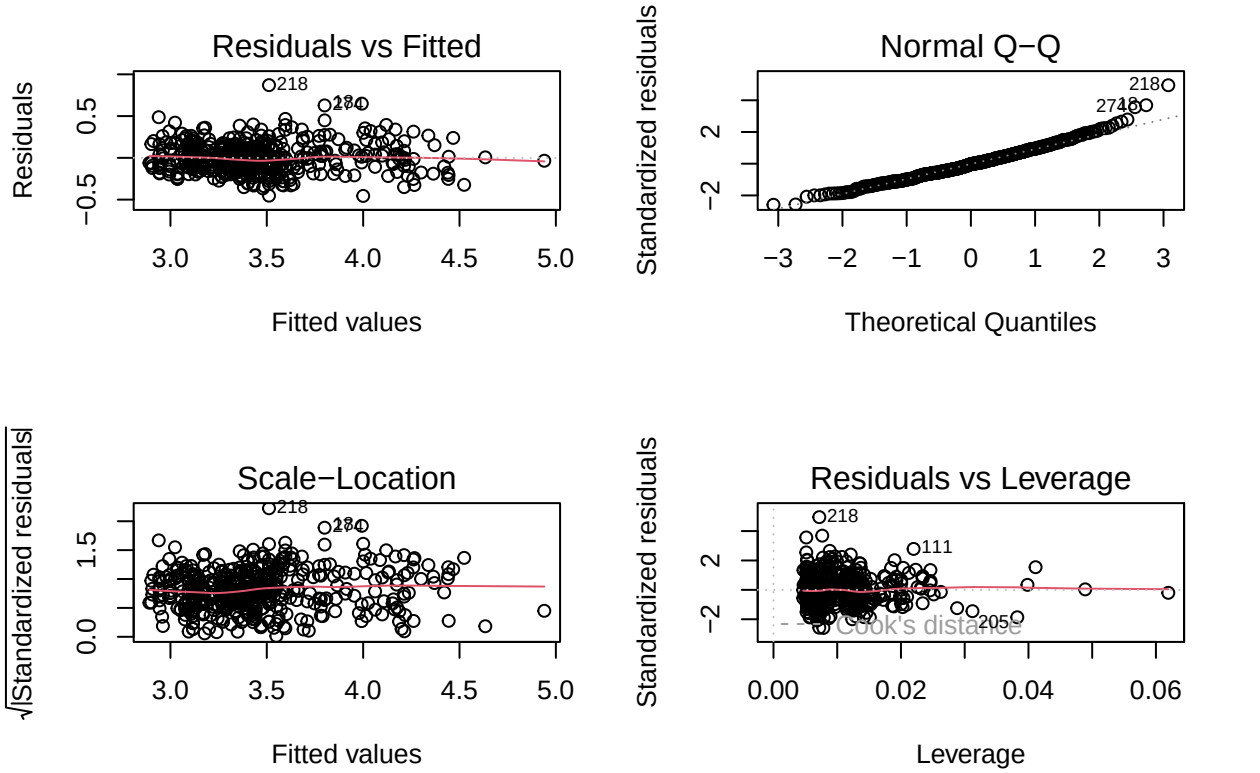
Alle estimater er signifikante på 0,1% bortset fra "Minority", som er signifikant på 1%.

F-testen afvises grundet den lave p-værdi angivet ved < 0.000000000000000222, hvilket vil sige at variablene er "jointly significant", da nulhypotesen i F-testen, $H_0 = \beta_{1,2,3,4} = 0$ afvises. Under antagelse af, at alle andre variable er faste, vil en stigning i "educ" (uddannelse) på 1, medfører en stigning i lønnen på 3,5%. Et ekstra års uddannelse vil altså øge lønnen med 3,5%. Samme princip er gældende for de tre andre variable "logbegin", "male", "minority". Derudover angiver ovenstående model at $R^2 = 0.77$, hvilket vil sige at dataen passer relativt godt til den estimerede model.

Opgave 2 - Udfør grafisk modelkontrol

For at udføre grafisk kontrol opstilles en række plots som beskrives nedenfor.

```
par(mfrow=c(2,2))
plot(model1)
```



Ovenstående er udført grafisk modelkontrol. "Residuals vs. Fitted" viser, at residualerne ikke er spredte, hvilket er indikation på, at der ikke er tale om et "non-linear relationship". Dog er den røde linje her næsten vandret (med undtagelse af det yderste af x-aksen), hvilket kunne være tegn homoskedasticitet og dermed være tegn på et lineært forhold. "Q-Q plot" viser at residualerne tilnærmelsesvist følger en ret linje, og derfor antages de at være normalfordelt. "Scale-Location" belyser, at punkterne ikke er vilkårligt fordelt, hvilket kan indikerer at der er tale om heteroskedasticitet for modellen. Ideelt ønskes, at den røde linje er vandret samt at residualpunkterne er spredt og tilfældigt fordelt. "Resudials vs. Leverage" tydeliggøre problematikken vedrørende outliers. Y-aksen angiver de standardiseret residualer givet ved $\frac{u}{se}$, mens x-aksen angiver "leverage" som måler hvor stor indflydelse en observation har på estimaterne for regressionens koefficienter. De stiplede linjer viser de forskellige niveauer for Cook's distance, som viser hvorvidt en outlier har stor betydning for estimationen af regressionens koefficienter. Hvis Cook's distancen for en observation er større end 1, anses den for at være indflydelsesrig for estimationen. I modellen ses tre outliers, 205, 343 og 29. Her ligger de to førstnævnte indenfor Cook's distance på 0,5 og vurderes derfir til at have lav indflydelse, mens 29 ligger på en Cook's distance imellem 0,5 og 1, hvilket heller ikke vurderes til at have den store effekt på regressionens estimater. Der er ingen outliers udenfor Cook's distance på 1.

Opgave 3 - Test for heteroskedasticitet vha. Breusch-Pagan-testen og specialudgaven af White-testen

For at teste for heteroskedasticitet udføres BP-testen samt specialudgaven af White-testen. Her udføres BP-testen både manuelt og vha. funktionen i R. BP-testen udføres ved at kvadrere fejlledet i den oprindelige regression, hvorefter denne opstilles som en funktion af de uafhængige variable fra den oprindelige regression.

FORMLER FOR BP

Der udføres en F-test eller LM-test for at estimere p-værdien, og hvis denne er under det valgte signifikansniveau afvises nulhypotesen.

$$H_0 : homoskedasticitet$$

Hvis p-værdien er lav i BP-testen/F-test vil H_0 afvises hvorfor der antages at være heteroskedasticitet.

White-testen udføres også ved at opstille en regression for det kvadreret fejlled. Dog opstilles denne med de uafængige variable, de kvadrerede uafhængige variable samt krydsprodukterne af de uafhængige variable. Igen udføres en F-test eller LM-test for at vurdere hvorvidt nulhypotesen afvises eller accepteres.

MANGLER TEKST OM SPECIALUDGAVE AF WHITE

```
u = resid(model1)
u2 = u^2
model1u = lm(u2 ~ educ+logbegin+male+minority) #Test for heteroskedasticitet
summary(model1u)
```

```
##
## Call:
## lm(formula = u2 ~ educ + logbegin + male + minority)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.047734 -0.025058 -0.013454  0.009077  0.717499
##
## Coefficients:
##               Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  0.00364042  0.02359924  0.1543   0.8775
## educ         0.00196987  0.00121572  1.6203   0.1058
## logbegin    -0.00080605  0.01132014 -0.0712   0.9433
## male         0.00948273  0.00625530  1.5160   0.1302
## minority    -0.01044968  0.00639085 -1.6351   0.1027
##
## Residual standard error: 0.055484 on 469 degrees of freedom
## Multiple R-squared:  0.029233,   Adjusted R-squared:  0.020954
## F-statistic: 3.5308 on 4 and 469 DF,  p-value: 0.0074747
```

Da nulhypotesen i F-testen afvises angiver det, at der er heteroskedasticitet i modellen. Ved at bruge χ^2 i stedet for F-test findes Breusch-Pagan testen.

```
lm_chi = 0.077789*474 #BP-test
1-pchisq(lm_chi, 4) #p-værdien for chi-square
```

```
## [1] 0.00000019140651
```

```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 13.8566, df = 4, p-value = 0.0077671
```

BP-værdien på 36.8719 indikerer, at variansen af residualerne ikke er konstant, hvilket også bekræftes ved den lave p-værdi for chi-square. Med den lave p-værdi, kan H_0 afvises, hvorfor det tyder på, at der findes heteroskedasticitet.

White-test med fitted værdier:\ Her benyttes residualer i anden som regresseres på de fitted værdier af modellen og de fitted værdier i anden for at belyse lineære og ikke-lineære forhold mellem de uafhængige variable og residualerne.

```
white = lm(u2 ~ predict(model1) + I(predict(model1)^2))
summary(white)
```

```
##
## Call:
## lm(formula = u2 ~ predict(model1) + I(predict(model1)^2))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.044319 -0.026041 -0.014620  0.007234  0.722793
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.306039   0.196031 -1.5612   0.1192
## predict(model1)       0.167264   0.108320  1.5442   0.1232
## I(predict(model1)^2) -0.019963   0.014833 -1.3458   0.1790
##
## Residual standard error: 0.055545 on 471 degrees of freedom
## Multiple R-squared:  0.022954,   Adjusted R-squared:  0.018805
## F-statistic: 5.5327 on 2 and 471 DF,  p-value: 0.0042168
```

```r
lm_chi = 0.022954*474 #BP-test
1-pchisq(lm_chi, 2) #p-værdien for chi-square
```

```
## [1] 0.004339058
```

```r
#bptest(model1, ~ predict(model1) + I(predict(model1)^2))
```

Igen afvises nulhypotesen i F-testen og der antages derfor at være heteroskedasticitet, hvilket også gælder når χ^2 bruges og dermed en BP-test. Både $\hat{y}$ og $\hat{y}^2$ er signifikante i testen, så det kan ikke siges hvorvidt der er et lineært eller ikke-lineært forhold.

Opgave 4 - Beregn robuste standardfejl for modellen og sammenlign med resultaterne i spørgsmål 1

MANGLER FORKLARING AF ROBUSTE SE

```r
model1robust <- coeftest(model1, vcov = vcovHC(model1, type = "HC0"))
screenreg(list(OLS = model1, OLS_robust_se = model1robust), digits = 4)
```

```
##
## ========================================
##              OLS           OLS_robust_se
## ----------------------------------------
## (Intercept)    0.8487 ***    0.8487 ***
##               (0.0751)      (0.0794)
## educ           0.0233 ***    0.0233 ***
##               (0.0039)      (0.0035)
## logbegin       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
```

Den venstre kolonne udgør resultaterne fra opgave 1, mens den højre kolonne viser samme estimater men med robuste standardafvigelser. Der findes en lille ændring i signifikansniveauet, hvor "minority" er signifikant på et højere niveau, hvilket skyldes den lavere standardafvigelse.

Opgave 5 - Test hypotesen H0: $\beta_2 = 1$ mod alternativet H1: $\beta_2 \neq 1$

MÅSKE NOGET OM T-FORDELING T-scoren beregnes med følgende formel

$$T = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$$

Her er $\hat{\beta}_j$ estimatet fra regressionen og β_j er nulhypotesen, som i dette tilfælde er 1. Der vil i udregningen blive brugt robust standardafvigelser grundet den påviste heteroskedasticitet\

```
#summary(model1)
t = (0.03008211-1) / 0.00296699
t
```

```
## [1] -326.90299
```

```
#kritiske værdier
alpha = c(0.05, 0.01)
qt(1-alpha/2, 469)
```

```
## [1] 1.9650350 2.5863526
```

```
#P-værdier
pt(-abs(t), 469)
```

```
## [1] 0
```

De kritiske værdier for 5% og og 1% er henholdsvis 1,96 og 2,59 hvorfor H_0 afvises og $\beta_2 \neq 1$. P-værdien rapporteres i R som 0, da t-scoren er så lav at p-værdien er for lav til at vise.

Opgave 6 - Test hypotesen H0: $\beta_3 = \beta_4 = 0$

F-test MANGLER FORKLARING OM F-TEST

```
linearHypothesis(model1, c("male=0", "minority=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## male = 0
## minority = 0
##
## Model 1: restricted model
```

```
## Model 2: logsal ~ educ + logbegin + male + minority
##
##   Res.Df     RSS Df Sum of Sq       F   Pr(>F)
## 1    471 14.8917
## 2    469 14.6275  2  0.264165 4.23495 0.015038 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#model1a = lm(logsal ~ educ+logbegin)
#waldtest(model1, model1a)
```

RSS (kan også kaldes SSR) beskriver residual sum of squares, som er summen af de estimeret fejlled sat i anden, og hvis SSR = 0 er modellen perfekt ($R^2 = 1$).

FORMEL FOR SSR

Den meget lave p-værdi, som tilnærmelsesvist er nul, gør at H0 kan afvises. Dette betyder, at estimaterne "male" eller "minority" er "jointly significant" og dermed i fællesskab forskellig fra 0.

For at beregne F-statistic, kan der gøres brug af følgende formel:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

Her angiver r den begrænsede model, hvor $\beta_3 = \beta_4 = 0$, mens ur angiver den ubegrænsede model hvor alle variable indgår. q er forskellen i frihedsgrader mellem de to modeller. $(n-k-1)$ er antal frihedsgrader i den ubegrænset model, hvor n er antal observationer og k er antal variable i modellen.\ Jf. ovenstående er F-værdien 22.6445, hvilket giver en lav p-værdi, hvorfor nulhypotesen, hvor β_3 og β_4 er lig med nul, afvises. Derfor vil mindst et af de to estimater have relevans for modellen.

Opgave 7 - Estimer modellen vha. FGLS og kommenter på resultaterne

MANGLER TEKST TIL FGLS

```
logu2 <- log(resid(model1)^2) #Her gøres alt i en command
varreg<-lm(logu2 ~ educ+logbegin+male+minority)
w <- exp(fitted(varreg))
model2fgls = lm(logsal ~ educ+logbegin+male+minority, weight=1/w)
summary(model2fgls)
```

```
##
## Call:
## lm(formula = logsal ~ educ + logbegin + male + minority, weights = 1/w)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -4.75399 -1.30933 -0.04817  1.15316  8.93169
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  0.8492577  0.0756272 11.2295 < 0.00000000000000022 ***
## educ         0.0221581  0.0037704  5.8769        0.000000007941 ***
## logbegin     0.8269565  0.0357798 23.1124 < 0.00000000000000022 ***
## male         0.0486547  0.0195960  2.4829               0.01338 *
## minority    -0.0428553  0.0186707 -2.2953               0.02216 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.948 on 469 degrees of freedom
## Multiple R-squared:  0.80456,    Adjusted R-squared:  0.8029
## F-statistic: 482.69 on 4 and 469 DF,  p-value: < 0.000000000000000222
```

Det ses hvordan alle koefficienter er statistisk signifikante på 0,1%. Desuden har modellen en høj forklaringsgrad på $R^2 = 0,73$. Alle estimaterne har relativt høje t-værdier, hvilket giver en lav p-værdi og dermed det høje signifikansniveau.

```
screenreg(list(OLS = model1, FGLS = model2fgls), digits = 4)
```

```
##
## ========================================
##              OLS           FGLS
## ----------------------------------------
## (Intercept)    0.8487 ***    0.8493 ***
##               (0.0751)      (0.0756)
## educ           0.0233 ***    0.0222 ***
##               (0.0039)      (0.0038)
## logbegin       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
```

MANGLER SAMMENLIGNING MED OLS

Opgave 8 - Har FGLS estimationen taget højde for al heteroskedasticiteten?

I ovenstående opgave 3 blev det tydeligt, at det signifikante resultat fra Breusch-Pagan testen indikerede, at modellen indeholdte heteroskedasticitet, hvilket er en god årsag til at benytte sig at FGLS estimationen. Denne tager nemlig højde for heteroskedasticitet i modellen, ved at omforme fejlledende så de bliver homoskedastiske (konstant varians). Anvendelsen af FGLS gør modellen mere præcis, end de resultater som opnås ved almindelig OLS, specielt når der er stærk heteroskedasticitet til stede. Det kan dog ikke udelukkes, at der efter FGLS estimationen ikke er mere heteroskedasticitet, og derfor kan der udeføres en BP-test eller White-test.

```
u2 = resid(model2fgls)^2
model2fgls_u = lm(u2 ~ educ+logbegin+male+minority, weight=1/w)
summary(model2fgls_u)
```

```
##
## Call:
## lm(formula = u2 ~ educ + logbegin + male + minority, weights = 1/w)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49937 -0.28288 -0.16006  0.09398  7.37572
##
## Coefficients:
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  0.00360176  0.02298331  0.1567  0.87554
## educ         0.00174549  0.00114584  1.5233  0.12835
## logbegin     0.00034174  0.01087357  0.0314  0.97494
## male         0.00952578  0.00595526  1.5996  0.11037
## minority    -0.01097277  0.00567407 -1.9338  0.05373 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.59199 on 469 degrees of freedom
## Multiple R-squared:  0.032353,   Adjusted R-squared:  0.0241
## F-statistic: 3.9202 on 4 and 469 DF,  p-value: 0.003841
```

Da nulhypotesen i F-testen afvises angiver det, at der stadig er heteroskedasticitet i modellen trods brug af FGLS. Ved at bruge χ^2 i stedet for F-test findes Breusch-Pagan testen.

```
lm_chi = 0.070322*474 #BP-test
1-pchisq(lm_chi, 4) #p-værdien for chi-square
```

```
## [1] 0.0000010210752
```

Det signifikante resultat fra Breusch-Pagan testen indikerer, at modellen stadig indeholder heteroskedasticitet.

```
white_fgls = lm(u2 ~ predict(model2fgls) + I(predict(model2fgls)^2))
summary(white_fgls)
```

```
##
## Call:
## lm(formula = u2 ~ predict(model2fgls) + I(predict(model2fgls)^2))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.044437 -0.025936 -0.014284  0.007326  0.725167
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -0.307674   0.196554 -1.5653   0.1182
## predict(model2fgls)        0.168008   0.108591  1.5472   0.1225
## I(predict(model2fgls)^2) -0.020040   0.014868 -1.3479   0.1783
##
## Residual standard error: 0.055657 on 471 degrees of freedom
## Multiple R-squared:  0.023107,   Adjusted R-squared:  0.018959
## F-statistic: 5.5704 on 2 and 471 DF,  p-value: 0.004064
```

White-testen viser ligeledes tegn på heteroskedasticitet, da p-værdien i F-testen er tilnærmelsesvis nul.

Eksamenssæt 2

Opgave 1 - Estimer de to modeller vha. OLS. Kommenter på outputtet, sammenlign og fortolk resultaterne.

```
model1 = lm(salary ~ educ + salbegin + male + minority, data = data2)
model2 = lm(log(salary) ~ educ + log(salbegin) + male + minority, data = data2)

screenreg(list(model1, model2), digits = 4)
```

```
##
## ========================================
##                Model 1       Model 2
## ----------------------------------------
## (Intercept)     -6.9323 ***    0.8491 ***
##                 (1.8554)      (0.0771)
## educ             0.9933 ***    0.0236 ***
##                 (0.1667)      (0.0040)
## salbegin         1.6082 ***
##                 (0.0641)
## male             1.8309 *      0.0455 *
##                 (0.8571)      (0.0208)
## minority        -1.7254       -0.0419 *
##                 (0.9206)      (0.0211)
## log(salbegin)                  0.8207 ***
##                               (0.0371)
## ----------------------------------------
## R^2              0.7962        0.8051
## Adj. R^2         0.7944        0.8034
## Num. obs.      450           450
## ========================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Model 1 og model 2 er forskellige fra hinanden idet variablen "salbegin" inddrages i model 1, mens variablen "log(salbegin)" er inddraget i model 2.

For model 1 er alle estimaterne signifikante med undtagelse af "minority", hvilket betyder at det ikke er sikkert at den har en betydning for den afhængige variabel. Skæringspunktet ("intercept"), "educ" og "salbegin" er signifikante på 0.1%, mens "male" er signifikant på 5%.

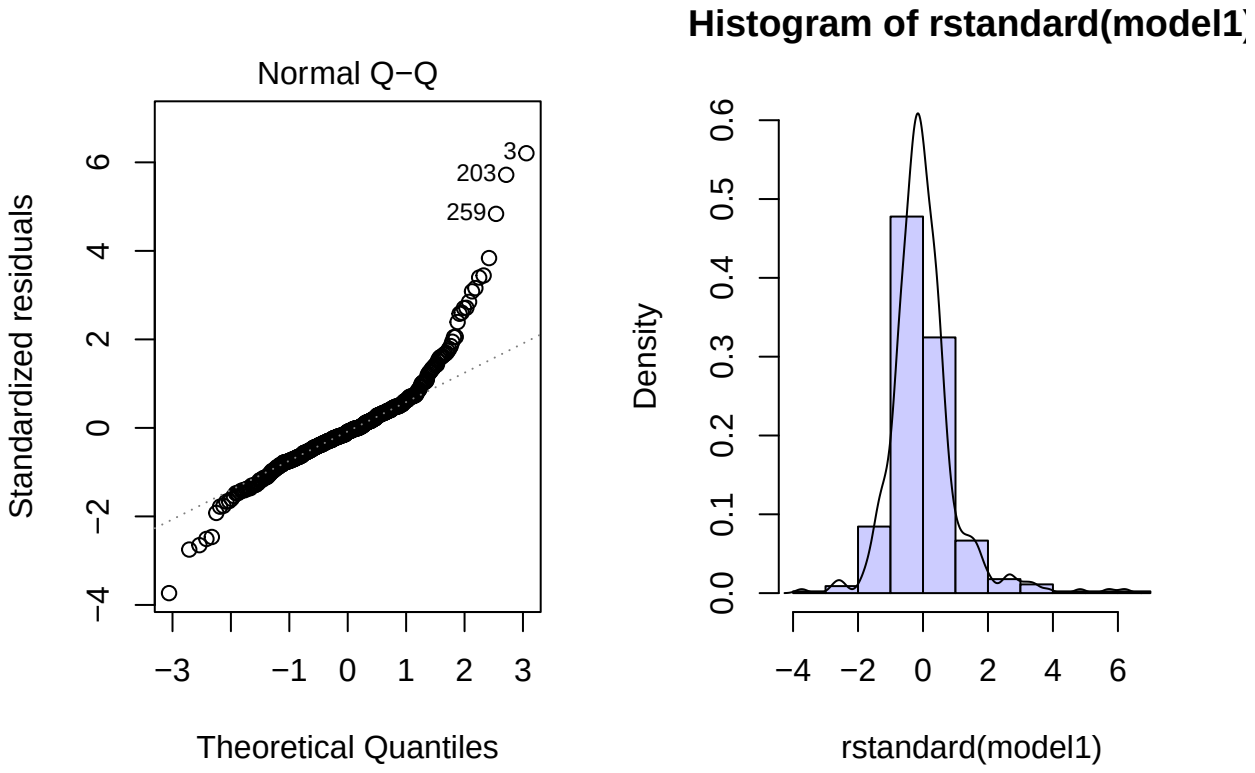
For model 2 er alle estimaterne signifikante, hvor skæringspunktet ("intercept"), "educ" og "log(salbegin)" er signifikante på 0.1%, mens estimaterne "male" og "minority" er signifikante på 5%.

Estimaterne i model 1 angiver, at en stigning i den enkelte variabel (under antagelse af, at de resterende variable holdes konstant) vil medføre en ændring i lønnen på det tilsvarende estimat.

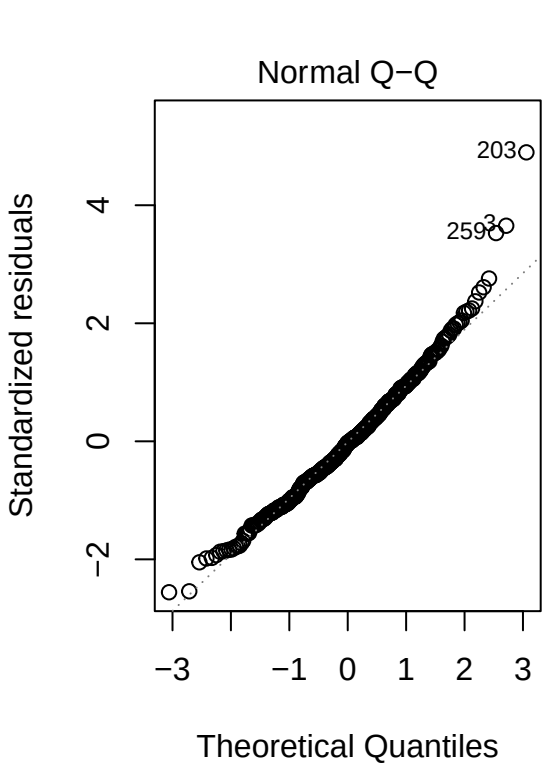
R^2 angiver forklaringsgraden af dataen i forhold til den estimeret model. Her kan det ses at $R^2 = 0.80$ i model 1, mens $R^2 = 0.81$ for model 2. Kigges der i stedet på "adjusted R^2", som også tager hensyn til hvorvidt flere variable tilføjes til modellen, kan det ses at model 1 har en $R^2 = 0.79$ mens model 2 har en $R^2 = 0.8$. Dermed uanset hvilken R^2 der tages udgangspunkt i passer dataen til model 2 1% bedre til modellen sammenlignet med model 1.

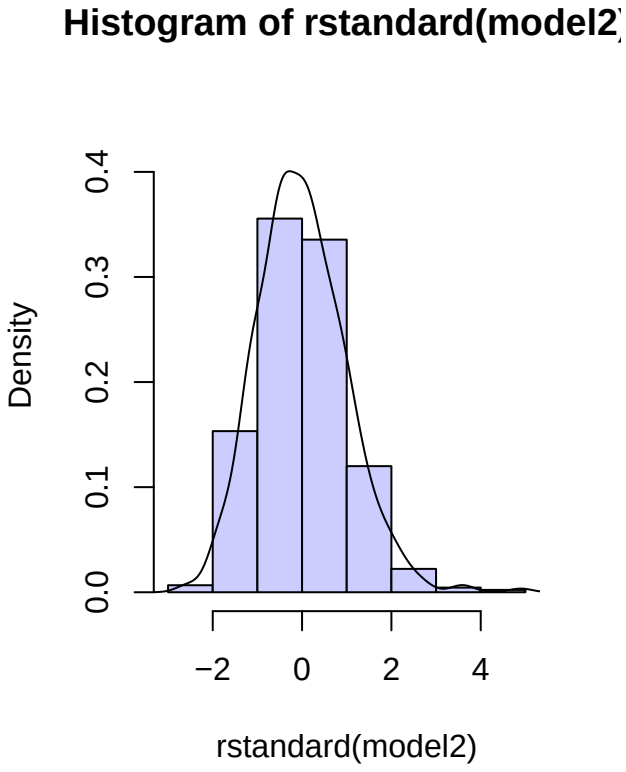
Opgave 2 - Udfør grafisk modelkontrol af de to modeller. Hvilken model vil du foretrække?

```
par(mfrow = c(1,2))
plot(model1, 2)
hist(rstandard(model1), prob=T, col = rgb(0.8,0.8,1), ylim=c(0,0.6))
lines(density(rstandard(model1)))
```

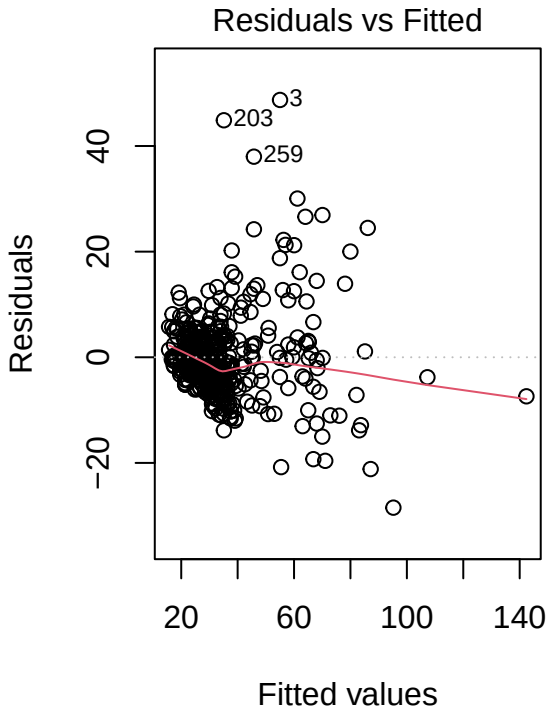


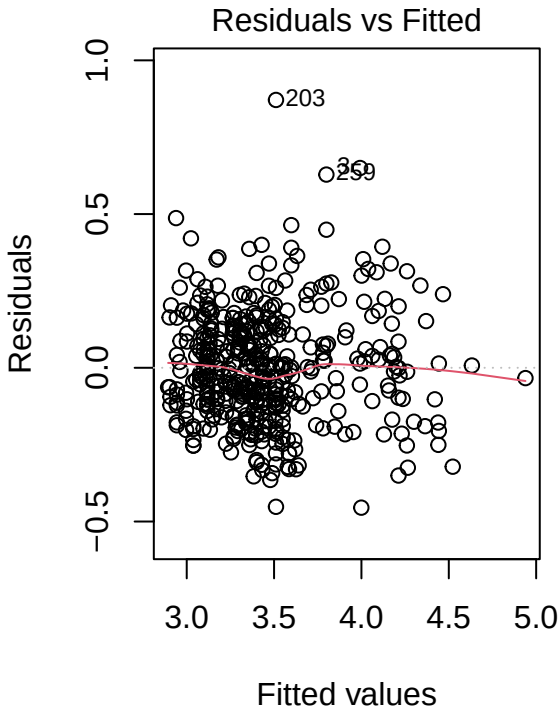

```
plot(model2, 2)
hist(rstandard(model2), prob=T, col = rgb(0.8,0.8,1), ylim=c(0,0.45))
lines(density(rstandard(model2)))
```






```
#Grafisk kontrol for misspecification
plot(model1, 1)
plot(model2, 1)
```





I ovenstående kan der ses tre plot for hver af de to modeller: "Q-Q residuals", historgram for de standard-

iserede residualer samt plot for "residuals vs. fitted"

For Q-Q residuals følger punkterne den stiplede rette linje. Hvis punkterne systematisk afviger fra den rette linje vil det tyde på, at residualerne ikke er normalfordelte. For model 1 passer punkterne til dels den rette linje. Dog er der også en del outliers især til højre for 0 på x-aksen. Modsat for model 2 passer de standardiserede residualer noget bedre til den rette linje. Dog er der en svag systematisk afvigelse i yderpunkterne, hvor punkterne ligger over den stiplede linje.

Histogrammet for standardiserede residualer vil være normalfordelt, hvis middelværdien for residualerne ligger omkring nul, hvilket vil indikere, at regressionens fejlled ikke over- eller undervurderer modellen systematisk. For model 1 ser fordelingen ud til at være en smule højreskæv, dog med en middelværdi på omkring 0. Samme tendens ses for model 2, hvor fordelingen her også er højreskæv, dog ligger middelværdien her en smule til venstre for nul.

For "Residuals vs. fitted" vil det ses, at punkterne er tilfældigt spredt over en vandret rød linje, hvis der er et lineært forhold mellem de uafhængige og afhængige variable. Residualerne for model 1 er ikke spredt, og den røde linje har en faldende tendens. Dog jo længere ud på x-aksen vi kommer jo mere tilfældigt spredte er de. Model 2 har samme tendens, hvor det dog ser ud til at punkterne er mere spredte (her er det vigtigt at være obs på x-aksens værdier, hvor x-aksen kun strækker sig fra 3-5 for model 2, mens den strækker sig fra 20-120 i model 1 - hvorfor det kan linje at model 1 er mindre spredt ift. model 2). Den røde linje for model 2 er dog fladere sammenlignet med model 1.

Opgave 3 - Undersøg om de to modeller er misspecificerede vha. RESET-testet

RESET står for "regression specification error test" og blev lavet af Ramsey i 1969. Hvis en model opfylder MLR4 (homoskedasticitet), kan man teste for misspecifikation, ved at tilføje den kvadrerede form af de uafhængige variable, og derefter udføre en F-test for at teste "joint null hypothesis". De kvadrerede variable vil fange, hvis der er noget ikke-lineært i modellen, og derfor vil de tilføjede kvadrerede variable være insignifikante for modellen, såfremt modellen ikke er misspecificeret.

```
y2 = fitted(model1)^2
y3 = fitted(model1)^3

reset1 = lm(salary ~ educ + salbegin + male + minority + y2 + y3, data = data2)

linearHypothesis(reset1, c("y2=0", "y3=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## y2 = 0
## y3 = 0
##
## Model 1: restricted model
## Model 2: salary ~ educ + salbegin + male + minority + y2 + y3
```

```
##
##   Res.Df   RSS Df Sum of Sq     F Pr(>F)
## 1    445 27600
## 2    443 27282  2     317.2 2.576 0.0773 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
resettest(model1)
```

```
##
##  RESET test
##
## data:  model1
## RESET = 2.576, df1 = 2, df2 = 443, p-value = 0.0773
```

```
y2 = fitted(model2)^2
y3 = fitted(model2)^3
reset2 = lm(log(salary) ~ educ + log(salbegin) + male + minority + y2 + y3, data = data2)
linearHypothesis(reset2, c("y2=0", "y3=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## y2 = 0
## y3 = 0
##
## Model 1: restricted model
## Model 2: log(salary) ~ educ + log(salbegin) + male + minority + y2 + y3
##
##   Res.Df   RSS Df Sum of Sq     F Pr(>F)
## 1    445 14.19
## 2    443 14.03  2    0.1668 2.634 0.0729 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
resettest(model2)
```

```
##
##  RESET test
##
## data:  model2
## RESET = 2.634, df1 = 2, df2 = 443, p-value = 0.0729
```

Begge er insignifikante på 5%, men signifikante på 10%. De er dermed misspecificeret på 10% men ikke 5%. Hvorvidt den er misspecificeret ud fra reset-test afhænger af signifikansniveauet.

Opgave 4 - Forklar hvorfor det kunne være relevant at medtage educ2 som forklarende variabel i de to modeller. Estimer de to modeller igen hvor educ2 inkluderes (med tilhørende koefficient ¯5), kommenter kort på outputtet og udfør RESET-testet igen.

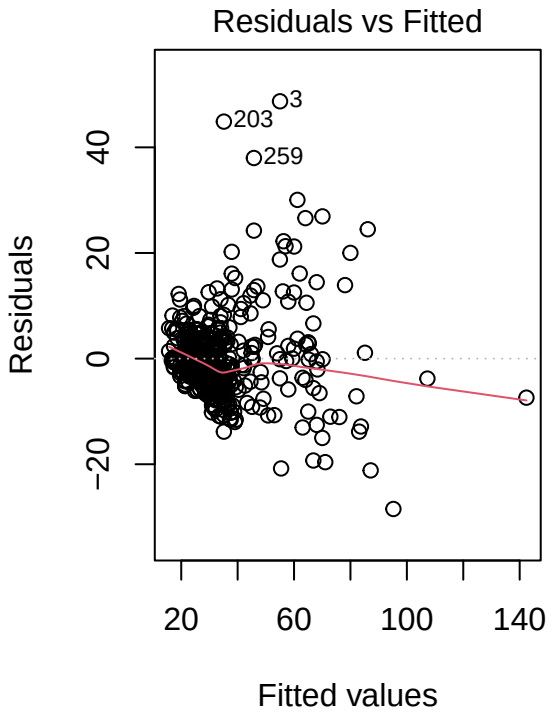
Som det fremgår af 2.3, kan modellen være misspecificeret. Det kan derfor efterprøves, om det skyldes en ikke-lineær sammenhæng mellem længden af uddannelse (educ) og løn, ved at inkludere educˆ2, for derefter lave RESET igen og se om der herefter vil være et klart svar på, hvorvidt modellen er mispecifiseret.

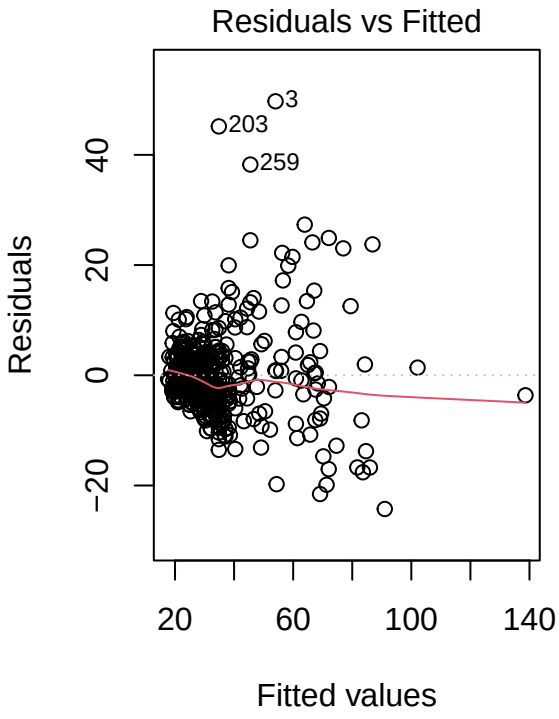
```
educ2 = data2$educˆ2
model1educ = lm(salary ~ educ + educ2 + salbegin + male + minority, data = data2)
model2educ = lm(log(salary) ~ educ + educ2 + log(salbegin) + male + minority, data = data2)

screenreg(list(model1educ, model2educ), digits = 4)
```

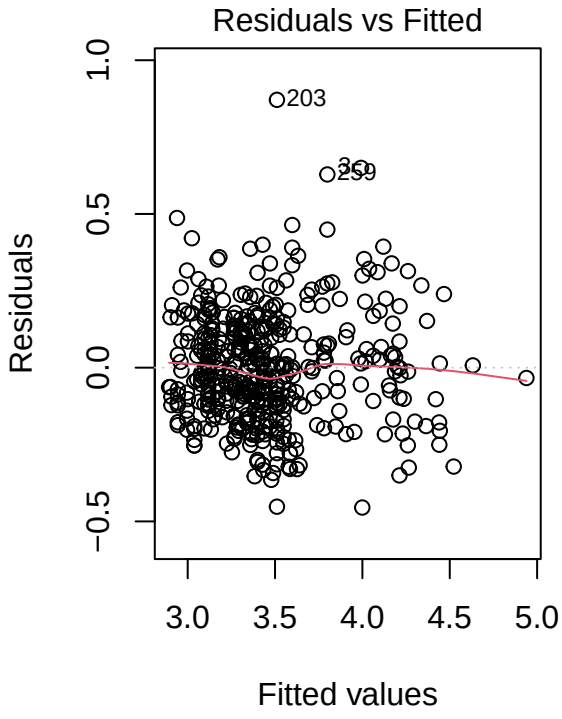
```
##
## ========================================
##                Model 1       Model 2
## ----------------------------------------
## (Intercept)     14.6024 *      1.1754 ***
##                 (6.7894)      (0.2073)
## educ            -2.3047 *     -0.0148
##                 (1.0146)      (0.0230)
## educ2            0.1321 **     0.0016
##                 (0.0401)      (0.0009)
## salbegin         1.4799 ***
##                 (0.0744)
## male             1.7855 *      0.0483 *
##                 (0.8479)      (0.0208)
## minority        -1.6150       -0.0416 *
##                 (0.9112)      (0.0210)
## log(salbegin)                  0.7825 ***
##                               (0.0433)
## ----------------------------------------
## Rˆ2              0.8011        0.8064
## Adj. Rˆ2         0.7989        0.8042
## Num. obs.      450           450
## ========================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

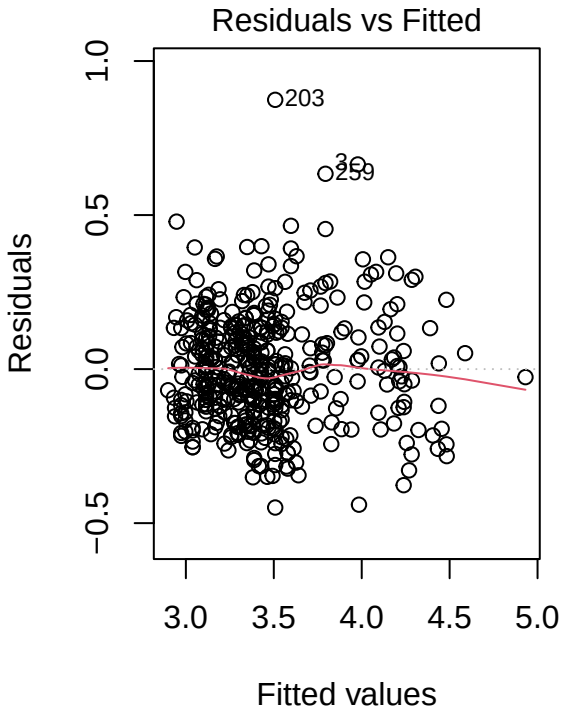
```
par(mfrow = c(1,2))
plot(model1, 1)
plot(model1educ, 1)
```






```
plot(model2, 1)
plot(model2educ, 1)
```






```
reset(model1educ)
```

```
##
```

```
##  RESET test
##
## data:  model1educ
## RESET = 1.877, df1 = 2, df2 = 442, p-value = 0.154
```

```
reset(model2educ)
```

```
##
##  RESET test
##
## data:  model2educ
## RESET = 1.888, df1 = 2, df2 = 442, p-value = 0.153
```

Der er højere p-værdier i begge, hvorfor de nu begge er insignifikante på 15% og under. Dermed er begge modeller nu ikke misspecificeret på noget relevant signifikansniveau. Dog er hverken *educ* eller $educ^2$ signifikante på 5% i model 2.

Opgave 5 - Test hypotesen $H0 : \beta_1 = \beta_5 = 0$ i begge modeller (fra spørgsmål 4).

MANGLER TEKST

```
linearHypothesis(model1educ, c("educ=0", "educ2=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ = 0
## educ2 = 0
##
## Model 1: restricted model
## Model 2: salary ~ educ + educ2 + salbegin + male + minority
##
##   Res.Df   RSS Df Sum of Sq     F         Pr(>F)
## 1    446 29801
## 2    444 26941  2      2860 23.56 0.000000000188 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#model1a = lm(logsal ~ educ+logbegin)
#waldtest(model1, model1a)

linearHypothesis(model2educ, c("educ=0", "educ2=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ = 0
## educ2 = 0
##
## Model 1: restricted model
## Model 2: log(salary) ~ educ + educ2 + log(salbegin) + male + minority
##
##   Res.Df   RSS Df Sum of Sq     F       Pr(>F)
## 1    446 15.31
## 2    444 14.10  2     1.203 18.94 0.0000000128 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#model1a = lm(logsal ~ educ+logbegin)
#waldtest(model1, model1a)
```

Lav p-værdi, så nulhypotese forkastes og β_1 og β_5 er derfor fælles signifikant

MANGLER FORTOLKNING

Opgave 6 - Kunne der være problemer med målefejl i de to modeller? I hvilke tilfælde vil det udgøre et problem?

Hvis den afhængige variabel, i dette tilfælde lønnen, har problemer med målefejl, vil dette komme til udtryk i fejlleddet i regressionen. Da lønnen, i tilfælde af målefejl, er den faktiske løn fratrukket dets fejlled, vil fejlleddet i regressionen være summen af det oprindelige fejlled, u, og målefejlen e_0. Det vil ikke skabe et problem så længe det oprindelige fejlled og målefejlen ikke er korreleret med nogle af de uafhængige variable, da det ellers vil skabe en bias i estimaterne.

Hvis den uafhængige variabel derimod har målefejl afhænger effekten af antagelserne vedrørende målefejlen. I dette tilfælde kunne der potentielt være en målefejl i variablen educ, da den kan dække over meget forskellig uddannelse som kan være svært at opgøre. Hvis der ikke er en korrelation mellem målefejlen og den observerede variabel, dvs. $cov(educ, e_{educ}) = 0$, hvor e_{educ} er målefejlen i variablen, er der nødvendigvis en korrelation mellem målefejlen og den ikke-observerede variabel $cov(educ^*, e_{educ}) \neq 0$. Derfor bliver fejlleddet i regressionen det oprindelige fejlled, u, fratrukket $\beta_1 e_{educ}$. I dette tilfælde vil estimatet af regressionen stadig være unbiased og consistent, da fejlleddet ikke er korreleret med den observerede variabel.

Hvis der er en korrelation mellem målefejlen og den observerede variabel $cov(educ, e_{educ}) \neq 0$, da opstår fejlen CEV (classical errors-in.variables). Dette betyder estimatoren er biased, og OLS altid vil undervurdere effekten af β. Da det er en multipel regression, vil en målefejl i dette tilfælde i en afhængig variabel betyde bias i alle andre parametre.

Eksamenssæt 3: Instrumentvariable

Opgave 1 - Estimer modellen vha. OLS og kommenter på resultaterne

```
model = lm(learnings ~ educ + exp + male + ethblack + ethhisp, data)
summary(model)
```

```
##
## Call:
## lm(formula = learnings ~ educ + exp + male + ethblack + ethhisp,
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.075849 -0.280064 -0.001448  0.307748  1.984409
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  0.3962265  0.1735078  2.2836             0.022801 *
## educ         0.1242201  0.0094515 13.1429 < 0.00000000000000022 ***
## exp          0.0338820  0.0050456  6.7152      0.00000000004986 ***
## male         0.2934491  0.0458032  6.4067      0.00000000033631 ***
## ethblack    -0.1956696  0.0712545 -2.7461             0.006243 **
## ethhisp     -0.0974063  0.1003417 -0.9707             0.332132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.51025 on 514 degrees of freedom
## Multiple R-squared:  0.35393,    Adjusted R-squared:  0.34764
## F-statistic: 56.316 on 5 and 514 DF,  p-value: < 0.000000000000000222
```

hispanics er insignifikant med p-værdi på 33,2% Resten er på 1% eller lavere Relativ lav R^2 F-test med meget lav p-værdi - Variable er "jointly significant"

Opgave 2 - Hvorfor kunne vi være bekymrede for at uddannelse er endogen?

Uddannelse vil være endogen hvis den er korreleret med en udeladt variabel som derfor skaber en bias. Denne udeladte variabel kunne eksempelvis være "ability", som påvirker uddannelsesniveauet positivt og dermed skaber en bias, eller forældrenes uddannelsesniveau. Ligeledes kan antallet af søskende have en negativ effekt på uddannelsesniveauet.

Opgave 3 - Er siblings, meduc og feduc brugbare som instrumenter?

Hvis de nævnte variable er korreleret med uddannelse, mens de ikke er korreleret med den udeladte variabel, som i dette tilfælde er "ability", vil de være egnet som instrumenter. Det kan formentlig antages, at forældre uddannelse eller antal søskende uddannelse ikke har indflydelse på "ability", hvorfor denne betingelse til instrumentvariablen er opfyldt. Samtidig er forældres uddannelsesniveau formentlig delvist korreleret med den pågældendes uddannelse, mens antal søskende ikke i samme grad antages at være korreleret med uddannelsesniveauet. Derfor vil forældres uddannelse formentlig være bedre instrumenter end antal søskende.

Opgave 4 - Test om uddannelse er endogen

Testen for endogenitet laves vha. den reduceret ligning, som er variablen mistænkt for endogenitetsproblemer regresseret på de øvrige uafhængige variable og instrumentvariablene. Heri er variablen eksogen, altså ukorreleret med det oprindelige fejlled (u), hvis og kun hvis fejlleddet fra den reducerede ligning (v) er ukorreleret med (u). Fejlleddet fra den reducerede ligning (v) er dog ikke observeret, hvorfor residualet bruges som proxy. Derfor inkluderes (v) i den oprindelige regresion, hvorefter en t-test bruges til teste hvorvidt den tilhørende estimator δ er signifikant. Hvis det findes, at δ ikke kan siges at være lig 0 er variablen endogen. Modsat vil variablen antages at være eksogen hvis nulhypotesen $H_0 : \gamma = 0$ ikke kan afvises.

```
#Reduced form equation
red_model = lm(educ ~ exp + male + ethblack + ethhisp + siblings + meduc + feduc, data)
v = resid(red_model)

#Test for endogenitet, hvor residualer fra ovenstående RFE er med. Signifkans af denne vil betyde endog
endo_model = lm(learnings ~ educ + exp + male + ethblack + ethhisp + v, data)
summary(endo_model)
```

```
##
## Call:
## lm(formula = learnings ~ educ + exp + male + ethblack + ethhisp +
##     v, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.121193 -0.279066 -0.003089  0.298275  2.074787
##
## 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 ***
## ethblack    -0.157544   0.075122 -2.0972            0.03647 *
```

```
## ethhisp     -0.069476   0.101739 -0.6829            0.49499
## v           -0.036550   0.023106 -1.5818            0.11430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.50951 on 513 degrees of freedom
## Multiple R-squared:  0.35706,    Adjusted R-squared:  0.34955
## F-statistic: 47.484 on 6 and 513 DF,  p-value: < 0.000000000000000222
```

V er insignifikant, så der er ikke endogenitet??? Hvis siblings fjernes er der endogenitet??? I'm confused

Opgave 5 - Estimer modellen vha. 2SLS hvor du gør brug af de tre beskrevne instrumenter. Sammenlign med resultaterne i spørgsmål 1.

MANGLER TEKST TIL 2SLS

```
educ_fitted = fitted(red_model)

linearHypothesis(red_model, c("meduc=0", "feduc=0", "siblings=0")) #Test at IVs er signifikante
```

```
## Linear hypothesis test
##
## Hypothesis:
## meduc = 0
## feduc = 0
## siblings = 0
##
## Model 1: restricted model
## Model 2: educ ~ exp + male + ethblack + ethhisp + siblings + meduc + feduc
##
##   Res.Df     RSS Df Sum of Sq       F                 Pr(>F)
## 1    515 2914.55
## 2    512 2297.79  3   616.752 45.8087 < 0.000000000000000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sls = lm(learnings ~ educ_fitted + exp + male + ethblack + ethhisp, data)
#summary(sls)
screenreg(list(OLS = model, two_SLS = sls), digits = 4)
```

```
##
## ========================================
##              OLS           two_SLS
```

```
## ---------------------------------------
## (Intercept)    0.3962 *     -0.0647
##               (0.1735)      (0.3762)
## educ           0.1242 ***
##               (0.0095)
## exp            0.0339 ***    0.0376 ***
##               (0.0050)      (0.0062)
## male           0.2934 ***    0.2905 ***
##               (0.0458)      (0.0508)
## ethblack      -0.1957 **    -0.1575
##               (0.0713)      (0.0834)
## ethhisp       -0.0974       -0.0695
##               (0.1003)      (0.1129)
## educ_fitted                  0.1530 ***
##                              (0.0228)
## ---------------------------------------
## R^2            0.3539        0.2065
## Adj. R^2       0.3476        0.1988
## Num. obs.    520           520
## =======================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```
#Nedenstående er 2SLS lavet i R
#sls_r = ivreg(learnings ~ educ + exp + male + ethblack + ethhisp | meduc + feduc + siblings + exp + ma
#summary(sls_r)
```

Opgave 6 - Udfør overidentifikationstestet. Hvad konkluderer du?

MANGLER. Lektion 13. IM STILL CONFUSED

Opgave 7 - Udfør hele analysen igen hvor du kun bruger meduc og feduc som instrumenter. Ændrer det på dine konklusioner?

```
#Reduced form equation
red_model = lm(educ ~ exp + male + ethblack + ethhisp + meduc + feduc, data)
v = resid(red_model)

#Test for endogenitet, hvor residualer fra ovenstående RFE er med. Signifkans af denne vil betyde endog
endo_model = lm(learnings ~ educ + exp + male + ethblack + ethhisp + v, data)
summary(endo_model)
```

```
##
```

```
## Call:
## lm(formula = learnings ~ educ + exp + male + ethblack + ethhisp +
##     v, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.129818 -0.275266 -0.003979  0.291163  2.069504
##
## 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 ***
## ethblack    -0.1491819  0.0752218 -1.9832            0.04787 *
## ethhisp     -0.0633502  0.1017049 -0.6229            0.53364
## v           -0.0441531  0.0233868 -1.8879            0.05960 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.50899 on 513 degrees of freedom
## Multiple R-squared:  0.35839,    Adjusted R-squared:  0.35088
## F-statistic: 47.758 on 6 and 513 DF,  p-value: < 0.000000000000000222
```

Der er endogenitet, fordi residualerne er signifikante (?).

```
educ_fitted = fitted(red_model)

linearHypothesis(red_model, c("meduc=0", "feduc=0")) #Test at IVs er signifikante
```

```
## Linear hypothesis test
##
## Hypothesis:
## meduc = 0
## feduc = 0
##
## Model 1: restricted model
## Model 2: educ ~ exp + male + ethblack + ethhisp + meduc + feduc
##
##   Res.Df     RSS Df Sum of Sq       F                Pr(>F)
## 1    515 2914.55
## 2    513 2319.33  2   595.217 65.8264 < 0.000000000000000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
sls2 = lm(learnings ~ educ_fitted + exp + male + ethblack + ethhisp, data)
#summary(sls)

#Sammenligning med før
screenreg(list(OLS = model, "Tre IVs" = sls, "To IVs" = sls2), digits = 4)
```

```
##
## ====================================================
##               OLS           Tre IVs       To IVs
## ----------------------------------------------------
## (Intercept)    0.3962 *     -0.0647       -0.1658
##               (0.1735)      (0.3762)      (0.3818)
## educ           0.1242 ***
##               (0.0095)
## exp            0.0339 ***    0.0376 ***    0.0384 ***
##               (0.0050)      (0.0062)      (0.0062)
## male           0.2934 ***    0.2905 ***    0.2898 ***
##               (0.0458)      (0.0508)      (0.0507)
## ethblack      -0.1957 **    -0.1575       -0.1492
##               (0.0713)      (0.0834)      (0.0834)
## ethhisp       -0.0974       -0.0695       -0.0634
##               (0.1003)      (0.1129)      (0.1128)
## educ_fitted                  0.1530 ***    0.1594 ***
##                              (0.0228)      (0.0231)
## ----------------------------------------------------
## R^2            0.3539        0.2065        0.2098
## Adj. R^2       0.3476        0.1988        0.2021
## Num. obs.    520           520           520
## ====================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```r
#Nedenstående er 2SLS lavet i R
#sls_r = ivreg(learnings ~ educ + exp + male + ethblack + ethhisp | meduc + feduc + exp + male + ethbla
#summary(sls_r)
```

Ingen store ændringer

Eksamenssæt 4

Opgave 1 - Opstil en lineær regressionsmodel for *participation* hvor du bruger de beskrevne forklarende variable.

(a) - Estimer modellen vha. OLS og kommenter på resultaterne.

MANGLER FORKLARING AF LPM + SVAGHEDER U ER PER DEFINITION HETEROSKEDASTIC - DER BRUGES ROBUST SE

```
model_ols = lm(participation ~ income + age + agesq + educ + youngkids + oldkids + foreign, data = data
robust_ols = coeftest(model_ols, vcov = vcovHC(model_ols, type = "HC0"))

screenreg(list(OLS = model_ols, OLS_robust_se = robust_ols), digits = 4)
```

```
##
## ========================================
##              OLS           OLS_robust_se
## ----------------------------------------
## (Intercept)   -0.3686      -0.3686
##               (0.2530)     (0.2358)
## income        -0.0035 ***  -0.0035 ***
##               (0.0007)     (0.0006)
## age            0.0634 ***   0.0634 ***
##               (0.0129)     (0.0119)
## agesq         -0.0009 ***  -0.0009 ***
##               (0.0002)     (0.0001)
## educ           0.0068       0.0068
##               (0.0060)     (0.0059)
## youngkids     -0.2390 ***  -0.2390 ***
##               (0.0314)     (0.0302)
## oldkids       -0.0475 **   -0.0475 **
##               (0.0172)     (0.0175)
## foreign        0.2572 ***   0.2572 ***
##               (0.0401)     (0.0401)
## ----------------------------------------
## R^2            0.1901
## Adj. R^2       0.1836
## Num. obs.    872
## ========================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```
#summary(model_ols)
```

Alle signifikante på 0,1% pånær oldkids på 1% og educ er ikke signifikant. INGEN FORSKEL PGA ROBUST SE

(b) - Test om den partielle effekt af uddannelse er forskellig fra nul.

For at teste hvorvidt den partielle effekt af en variabel er forskellig fra nul bruges en t-test. Hvorvidt nulhypotesen afvises afhænger af den beregnede t-score og dertilhørende p-værdi

$$H_0 : \beta_4 = 0$$

$$H_1 : \beta_4 \neq 0$$

T-scoren beregnes ud fra den estimerede β samt den tilhørende standardafvigelse. Dette kan gøres, da nulhypotesen er, at den faktiske værdi er nul, hvorfor dette led ikke indgår i formlen.

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

I nedenstående vil de robuste standardafvigelser blive benyttet til udregningen af t-scoren

```
## Kritisk værdi ved 5% =  1.9626913
```

```
## Kritisk værdi vec 1% =  2.5814857
```

```
t = 0.0068/0.0059
```

```
## t-score =  1.1525424
```

Da t-scoren er under den kritiske værdi kan H_0 ikke afvises. Det samme vil ses i beregningen af p-værdien nedenfor.

```
2*(1-pt(t, df=length(data$educ)-1))
```

```
## [1] 0.24941449
```

Da p-værdien er højere end det den fastsatte grænse på 5% er resultatet statistisk signifikant.

(c) - Test om den partielle effekt af alder er forskellig fra nul.

For at teste hvorvidt alder er

Opgave 2 - Opstil både en logit- og en probit-model for par t i cipat ion hvor du bruger de beskrevne forklarende variable.

(a) - Estimer modellerne.

```
screenreg(list("LPM OLS" = model_ols, Logit = logit, Probit = probit))
```

```
##
## ================================================
##                  LPM OLS     Logit        Probit
## ------------------------------------------------
## (Intercept)      -0.37        -4.39 ***    -2.67 ***
##                  (0.25)       (1.30)       (0.78)
## income           -0.00 ***    -0.02 ***    -0.01 ***
##                  (0.00)       (0.00)       (0.00)
## age               0.06 ***     0.33 ***     0.20 ***
##                  (0.01)       (0.07)       (0.04)
## agesq            -0.00 ***    -0.00 ***    -0.00 ***
##                  (0.00)       (0.00)       (0.00)
## educ              0.01         0.04         0.02
##                  (0.01)       (0.03)       (0.02)
## youngkids        -0.24 ***    -1.18 ***    -0.71 ***
##                  (0.03)       (0.17)       (0.10)
## oldkids          -0.05 **     -0.24 **     -0.14 **
##                  (0.02)       (0.08)       (0.05)
## foreign           0.26 ***     1.19 ***     0.73 ***
##                  (0.04)       (0.20)       (0.12)
## ------------------------------------------------
## R^2               0.19
## Adj. R^2          0.18
## Num. obs.       872          872          872
## AIC                          1032.15      1031.65
## BIC                          1070.32      1069.82
## Log Likelihood               -508.08      -507.83
## Deviance                     1016.15      1015.65
## ================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

ESTIMATER FOR LOGIT OG PROBIT KAN IKKE FORTOLKES SOM DE ER

(b) - Test om den partielle effekt af uddannelse er forskellig fra nul.

(c) - Test om den partielle effekt af alder er forskellig fra nul vha. et likelihoodratio-test.

Opgave 3 - Vi vil gerne sammenligne den partielle effekt af *income* på tværs af modellerne. Beregn average partial effect (APE) og kommenter på resultaterne.

BRUGER ROBUST SE

```
ape_logit = logitmfx(logit, data = data, atmean=F, robust = T)

screenreg(list(ape_logit = ape_logit), digits = 4)
```

```
##
## ============================
##                  ape_logit
## ----------------------------
## income            -0.0046 ***
##                   (0.0010)
## age                0.0657 ***
##                   (0.0139)
## agesq             -0.0009 ***
##                   (0.0002)
## educ               0.0077
##                   (0.0060)
## youngkids         -0.2350 ***
##                   (0.0403)
## oldkids           -0.0470 **
##                   (0.0176)
## foreign            0.2466 ***
##                   (0.0409)
## ----------------------------
## Num. obs.          872
## Log Likelihood  -508.0766
## Deviance        1016.1533
## AIC             1032.1533
## BIC             1070.3196
## ============================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

```
ape_logit
```

```
## Call:
```

```
## logitmfx(formula = logit, data = data, atmean = F, robust = T)
##
## Marginal Effects:
##                  dF/dx    Std. Err.        z           P>|z|
## income    -0.004610926  0.001012942 -4.55201 0.0000053135052 ***
## age        0.065744335  0.013887098  4.73420 0.0000021991793 ***
## agesq     -0.000929881  0.000175837 -5.28832 0.0000001234475 ***
## educ       0.007705869  0.006036510  1.27654       0.2017634
## youngkids -0.235006974  0.040327957 -5.82740 0.0000000056299 ***
## oldkids   -0.046973245  0.017550880 -2.67640       0.0074417 **
## foreign    0.246583549  0.040923752  6.02544 0.0000000016865 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "foreign"
```

Opgave 4 - Vi vil gerne sammenligne den partielle effekt af *foreign* på tværs af modellerne. Beregn APE og kommenter på resultaterne.

Opgave 5 - Hvorfor er APE at foretrække frem for partial effect at the average (PEA)?

Opgave 6 - Sammenlign modellernes evne til at prædiktere ved at beregne percent correctly predicted for hver model.