Assignment 3, MSB104-gruppe 6 2025

Karin Liang

Introduksjon

Denne oppgaven har som mål å utforske alternative funksjonelle former og bruke panelestimater for å analysere forholdet mellom regional utvikling og inntektsulikhet i Belgia, Bulgaria, Nederland og Norge.

```
# Pakker
library(dplyr)
library(dineq)
library(tidyr)
library(ggplot2)
library(broom)
library(eurostat)
library(readxl)
library(lmtest)     # For bptest og coeftest
library(car)        # For ncvTest (score-test)
library(tibble)     # For å lage tabeller
library(fixest)
library(plm)
library(modelsummary)
```

Del A: Testing av utviklingseffekter på tvers av delmengder

Delmengdeanalyse

Resultatene fra del B i oppgave 2 viser at utdanning er den mest signifikante faktoren i den multiple lineære regresjonsmodellen. Derfor velger jeg å bruke utdanning som grunnlag for å teste delutvalg (subset) i denne analysen.

Datainnsamling

```
# Laster ned utdanningsdatasettet fra Eurostat
utdanning <- get_eurostat("edat_lfse_04", time_format = "num")
```

```
indexed 0B in  0s, 0B/s
indexed 1.00TB in  0s, 47.34TB/s
```

```r
# Filtrerer utdanningsdata for 2017 og grupperer regionene etter lav, middels og høy utdanni
utdanning_2017 <- utdanning %>%
  filter(
    TIME_PERIOD == 2017,
    sex == "T",             # Begge kjønn
    age == "Y25-64",        # Aldersgruppen 25-64 år
    unit == "PC",           # Enhet: prosent av befolkningen
    nchar(geo) == 4,        # Beholder kun NUTS2-nivå (fire tegn)
    substr(geo, 1, 2) %in% c("BE","NL","BG","NO"),
    isced11 != "ED3-8"
  )%>%
  mutate(edu_group = case_when(
    isced11 == "ED0-2" ~ "Low",
    isced11 %in% c("ED3_4", "ED3_4GEN", "ED3_4VOC") ~ "Medium",
    isced11 == "ED5-8" ~ "High",
    TRUE ~ NA_character_
  )) # lager tre utdanningsnivå (lav, middels, høy)
# Gjør utdanningsdata om til bredt format med tre kolonner for lav, middels og høy utdanning
utdanning_2017_trenivå <- utdanning_2017 %>%
  select(geo, edu_group, values) %>%
  pivot_wider(names_from = edu_group, values_from = values) %>%
  rename(
    `Lav_utdanning`     = Low,
    `Middels_utdanning` = Medium,
    `Høy_utdanning`     = High
  )
```

```r
# leser excel data_2017_b som brukt i assignment 2 del b
data_2017_b <- read_excel("/Users/liang/Desktop/msb104/msb104_a3/data_2017_b.xlsx", sheet =
# Slå sammen med hoveddatasettet (NUTS2 <-> geo)
data_2017 <- data_2017_b %>%
  left_join(utdanning_2017_trenivå, by = c("NUTS2" = "geo"))%>%
  select(-utdanning_hoy)
```

Estimering av effekten av tre utdanningsnivåer på ulikhet

```r
# Low education subset
model_lav  <- lm(Gini_weighted ~ change_GDPC_pct + vei_tetthet + andelen_over65 + Lav_utdann
summary(model_lav)
```

```
Call:
lm(formula = Gini_weighted ~ change_GDPC_pct + vei_tetthet +
    andelen_over65 + Lav_utdanning, data = data_2017)

Residuals:
      Min        1Q    Median        3Q       Max
-0.060821 -0.023907 -0.009843  0.027464  0.068873

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.2625760  0.1001004   2.623   0.0178 *
change_GDPC_pct  0.0033063  0.0056751   0.583   0.5678
vei_tetthet     -0.0006370  0.0004469  -1.425   0.1722
andelen_over65  -0.0031710  0.0047134  -0.673   0.5101
Lav_utdanning   -0.0045657  0.0021213  -2.152   0.0460 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04106 on 17 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5122,    Adjusted R-squared:  0.3975
F-statistic: 4.463 on 4 and 17 DF,  p-value: 0.01198
```

```r
# Medium education subset
model_middels  <- lm(Gini_weighted ~ change_GDPC_pct + vei_tetthet + andelen_over65 + Middel
summary(model_middels)
```

```
Call:
lm(formula = Gini_weighted ~ change_GDPC_pct + vei_tetthet +
    andelen_over65 + Middels_utdanning, data = data_2017)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06053 -0.02096 -0.01204  0.01644  0.10939

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.3705031  0.1431933   2.587   0.0192 *
change_GDPC_pct   0.0135135  0.0077165   1.751   0.0979 .
vei_tetthet      -0.0010718  0.0004906  -2.185   0.0432 *
andelen_over65   -0.0049287  0.0048401  -1.018   0.3228
Middels_utdanning -0.0044482  0.0027509  -1.617   0.1243
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.04312 on 17 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4621,    Adjusted R-squared:  0.3355
F-statistic: 3.651 on 4 and 17 DF,  p-value: 0.02534
```

```r
# High education subset
model_hoy  <- lm(Gini_weighted ~ change_GDPC_pct + vei_tetthet + andelen_over65 + Høy_utdann
summary(model_hoy)
```

```

Call:
lm(formula = Gini_weighted ~ change_GDPC_pct + vei_tetthet +
    andelen_over65 + Høy_utdanning, data = data_2017)

Residuals:
      Min        1Q    Median        3Q       Max
-0.056288 -0.021771 -0.008074  0.018879  0.067027

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.0141580  0.1171323  -0.121  0.90521
change_GDPC_pct  0.0112649  0.0055265   2.038  0.05738 .
vei_tetthet     -0.0009082  0.0004069  -2.232  0.03934 *
andelen_over65  -0.0028665  0.0043247  -0.663  0.51634
Høy_utdanning    0.0043338  0.0014869   2.915  0.00966 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03782 on 17 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5861,    Adjusted R-squared:  0.4888
F-statistic: 6.019 on 4 and 17 DF,  p-value: 0.003313
```

```r
modelsummary(
  list(
    "Low education regions"    = model_lav,
    "Medium education regions" = model_middels,
    "High education regions"   = model_hoy
  ),
  stars = TRUE,
  statistic = "({statistic})",
  fmt = 5,
  output = "markdown"
  )
```

Table 1: Resultater av delmengderegresjon på tvers av utdanningsgrupper

	Low education regions	Medium education regions	High education regions
(Intercept)	0.26258*	0.37050*	-0.01416
	(2.62313)	(2.58743)	(-0.12087)
change_GDPC_pct	0.00331	0.01351+	0.01126+
	(0.58259)	(1.75123)	(2.03835)
vei_tetthet	-0.00064	-0.00107*	-0.00091*
	(-1.42518)	(-2.18457)	(-2.23227)
andelen_over65	-0.00317	-0.00493	-0.00287
	(-0.67277)	(-1.01832)	(-0.66282)
Lav_utdanning	-0.00457*		
	(-2.15233)		
Middels_utdanning		-0.00445	
		(-1.61701)	
Høy_utdanning			0.00433**
			(2.91460)
Num.Obs.	22	22	22
R2	0.512	0.462	0.586
R2 Adj.	0.397	0.335	0.489
AIC	-71.7	-69.6	-75.3
BIC	-65.2	-63.0	-68.8
Log.Lik.	41.859	40.782	43.666
F	4.463	3.651	6.019
RMSE	0.04	0.04	0.03

- $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Diskusjon om delmengdeanalyse: (200–400 ord)

Som vist i tabell Table 1 undersøker delmengdeanalysen om sammenhengen mellom regional økonomisk utvikling og inntektsulikhet varierer på tvers av regioner med ulike utdanningsprofiler. Når de andre tre variablene holdes konstante, viser resultatene at ulike utdanningsnivåer som lavt, middels og høyt påvirker inntektsulikheten på forskjellige måter. I regioner med lav utdanning er utviklingseffekten (change_GDPC_pct) positiv, men ikke signifikant (omtrent estimate 0.003 og p 0.567), noe som tyder på at økonomisk vekst alene ikke reduserer ulikheten. Lav utdanning har en negativ og signifikant effekt (omtrent estimate -0.0046 og p 0.046), som betyr at økt utdanningsnivå i lavt utdannede regioner bidrar til å redusere forskjeller. Her spiller utdanning en sterkere rolle enn både økonomisk utvikling og infrastruktur. I regioner med middels utdanning er utviklingseffekten svak positiv (omtrent p 0.098), mens middels utdanning har en negativ, men ikke signifikant effekt (p 0.124). Veitetthet viser derimot en tydelig negativ og signifikant effekt ($p = 0.043$). Dette indikerer at infrastruktur har større forklaringskraft enn utdanning i disse områdene, og at bedre tilkobling mellom regioner kan bidra mer effektivt til å redusere inntektsulikhet enn økt utdanningsnivå. I regioner med høy utdanning er både økonomisk utvikling ($p = 0.057$) og høy utdanning ($p = 0.0097$) positivt

og signifikant knyttet til ulikhet. Veitetthet har derimot en negativ og signifikant effekt (estimate −0.0009, p 0.039), noe som tyder på at bedre infrastruktur kan bidra til å dempe forskjeller som ellers øker med økonomisk vekst og høyt utdanningsnivå. Dette antyder at når utdanningsnivået allerede er høyt, fører videre vekst og utdanning til at inntektsforskjellene øker, sannsynligvis fordi gevinsten av utviklingen tilfaller de mest kvalifiserte gruppene. Samlet viser analysen at utdanning reduserer ulikhet i lavt utdannede regioner, men kan forsterke forskjeller i høyt utdannede regioner, mens infrastruktur fremstår som en mer stabil og gjennomgående utjevnende faktor.

Del B: Utforsking av alternative funksjonsformer

Utforsk alternative funksjonsformer

```
# Log-lineær modell
model_log <- lm(Gini_weighted ~ log(change_GDPC_pct + 1) + vei_tetthet + andelen_over65 + `H
summary(model_log)
```

```
Call:
lm(formula = Gini_weighted ~ log(change_GDPC_pct + 1) + vei_tetthet +
    andelen_over65 + Høy_utdanning, data = data_2017)

Residuals:
      Min        1Q    Median        3Q       Max
-0.059703 -0.023982 -0.007641  0.019076  0.063994

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -0.1215117  0.1296854  -0.937  0.36189
log(change_GDPC_pct + 1)   0.0826641  0.0349689   2.364  0.03025 *
vei_tetthet               -0.0008725  0.0003881  -2.248  0.03813 *
andelen_over65            -0.0023623  0.0040748  -0.580  0.56970
Høy_utdanning              0.0046663  0.0014714   3.171  0.00558 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0366 on 17 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6124,    Adjusted R-squared:  0.5212
F-statistic: 6.715 on 4 and 17 DF,  p-value: 0.001968
```

```
# Kvadratisk modell
model_kvad <- lm(Gini_weighted ~ change_GDPC_pct + I(change_GDPC_pct^2) + vei_tetthet + ande
summary(model_kvad)
```

```
Call:
lm(formula = Gini_weighted ~ change_GDPC_pct + I(change_GDPC_pct^2) +
    vei_tetthet + andelen_over65 + Høy_utdanning, data = data_2017)

Residuals:
      Min        1Q    Median        3Q       Max
-0.058387 -0.026599 -0.003918  0.020466  0.057709

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.1436316  0.1400968  -1.025  0.32051
change_GDPC_pct       0.0369684  0.0173900   2.126  0.04943 *
I(change_GDPC_pct^2) -0.0019996  0.0012882  -1.552  0.14016
vei_tetthet          -0.0008640  0.0003920  -2.204  0.04253 *
andelen_over65       -0.0004593  0.0044358  -0.104  0.91881
Høy_utdanning         0.0047546  0.0014544   3.269  0.00482 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03635 on 16 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6403,    Adjusted R-squared:  0.5279
F-statistic: 5.696 on 5 and 16 DF,  p-value: 0.003328
```

```r
# Kubisk modell
model_kub <- lm(Gini_weighted ~ change_GDPC_pct + I(change_GDPC_pct^2) + I(change_GDPC_pct^3
summary(model_kub)
```

```
Call:
lm(formula = Gini_weighted ~ change_GDPC_pct + I(change_GDPC_pct^2) +
    I(change_GDPC_pct^3) + vei_tetthet + andelen_over65 + Høy_utdanning,
    data = data_2017)

Residuals:
      Min        1Q    Median        3Q       Max
-0.049406 -0.023702 -0.002343  0.015532  0.054563

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -7.947e-02  1.698e-01  -0.468  0.64652
change_GDPC_pct      -1.539e-03  5.824e-02  -0.026  0.97926
I(change_GDPC_pct^2)  4.028e-03  8.785e-03   0.459  0.65313
I(change_GDPC_pct^3) -2.800e-04  4.034e-04  -0.694  0.49834
vei_tetthet          -7.795e-04  4.167e-04  -1.871  0.08104 .
andelen_over65       -2.611e-05  4.552e-03  -0.006  0.99550
```

```
Høy_utdanning          4.502e-03  1.523e-03   2.957  0.00979 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03695 on 15 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.6515,    Adjusted R-squared:  0.5121
F-statistic: 4.673 on 6 and 15 DF,  p-value: 0.007158
```

```r
modelsummary(
  list(
    "Log-lineær modell" = model_log,
    "Kvadratisk modell" = model_kvad,
    "Kubisk modell"     = model_kub
  ),
  stars = TRUE,
  statistic = "({statistic})",   # viser t-verdier i parentes
  fmt = 5,
  output = "markdown"
)
```

Table 2: logaritmiske, kvadratiske, kubiske

	Log-lineær modell	Kvadratisk modell	Kubisk modell
(Intercept)	-0.12151	-0.14363	-0.07947
	(-0.93697)	(-1.02523)	(-0.46799)
log(change_GDPC_pct + 1)	0.08266*		
	(2.36393)		
vei_tetthet	-0.00087*	-0.00086*	-0.00078+
	(-2.24811)	(-2.20387)	(-1.87064)
andelen_over65	-0.00236	-0.00046	-0.00003
	(-0.57973)	(-0.10355)	(-0.00574)
Høy_utdanning	0.00467**	0.00475**	0.00450**
	(3.17124)	(3.26921)	(2.95712)
change_GDPC_pct		0.03697*	-0.00154
		(2.12585)	(-0.02643)
I(change_GDPC_pct^2)		-0.00200	0.00403
		(-1.55222)	(0.45855)
I(change_GDPC_pct^3)			-0.00028
			(-0.69392)
Num.Obs.	22	22	22
R2	0.612	0.640	0.651
R2 Adj.	0.521	0.528	0.512
AIC	-76.8	-76.4	-75.1
BIC	-70.2	-68.8	-66.4

Table 2: logaritmiske, kvadratiske, kubiske

	Log-lineær modell	Kvadratisk modell	Kubisk modell
Log.Lik.	44.387	45.209	45.557
F	6.715	5.696	4.673
RMSE	0.03	0.03	0.03

- p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Begrunn valgene mine

Blant de tre funksjonelle formene gir både den log-lineære og den kvadratiske modellen en bedre tilpasning enn den lineære modellen. Den log-lineære modellen har lavest AIC (−76.77), mens den kvadratiske har høyest justert R^2 (0.528), slik resultatene i tabell Table 2 viser. Jo høyere R^2, desto bedre; jo lavere AIC, desto bedre. Begge modellene gir dermed en bedre empirisk forklaring på sammenhengen mellom økonomisk utvikling og ulikhet, noe som også støttes teoretisk av Kuznets-hypotesen Kuznets (2019), der ulikheten først øker og deretter reduseres med økonomisk utvikling. Resultatene støttes av artikkelen Lessmann and Seidel (2017), om fant at ulikhet først øker og deretter avtar med økonomisk utvikling, og viser en N-formet sammenheng, noe som tilsvarer den klassiske Kuznets-kurven Kuznets (2019).

Visualisering

```
# Lager et nytt datasett for visualisering av modellene
visualisering_data <- data.frame(
  change_GDPC_pct = seq(min(data_2017$change_GDPC_pct, na.rm = TRUE),
                        max(data_2017$change_GDPC_pct, na.rm = TRUE),
                        length.out = 100), # Oppretter 100 jevnt fordelte verdier for økonom
  vei_tetthet = mean(data_2017$vei_tetthet, na.rm = TRUE),
  andelen_over65 = mean(data_2017$andelen_over65, na.rm = TRUE),
  Høy_utdanning = mean(data_2017$`Høy_utdanning`, na.rm = TRUE)) # Holder de andre forklarin

# Beregner predikerte Gini-verdier fra den log-lineære modellen
visualisering_data$pred_log  <- predict(model_log,  newdata = visualisering_data)
# Beregner predikerte Gini-verdier fra den kvadratiske modellen
visualisering_data$pred_kvad <- predict(model_kvad, newdata = visualisering_data)
```

```
visualisering <- data_2017 %>%
  ggplot(aes(x = change_GDPC_pct, y = Gini_weighted)) +
  geom_point(alpha = 0.6, color = "grey40") +
  geom_line(data = visualisering_data, aes(y = pred_log, color = "Log-lineær modell"), size
  geom_line(data = visualisering_data, aes(y = pred_kvad, color = "Kvadratisk modell"), size
  scale_color_manual(values = c("Log-lineær modell" = "blue", "Kvadratisk modell" = "red"))
  labs(
    title = "Sammenheng mellom økonomisk utvikling og inntektsulikhet",
```

```
    subtitle = "Sammenligning av log-lineær og kvadratisk modell",
    x = "Endring i BNP per innbygger (%)",
    y = "Gini-koeffisient",
    color = "Modell"
  ) +
  theme_minimal(base_size = 13)  +
  theme(
    legend.text = element_text(size = 5),
    legend.title = element_text(size = 6)
  )
visualisering
```

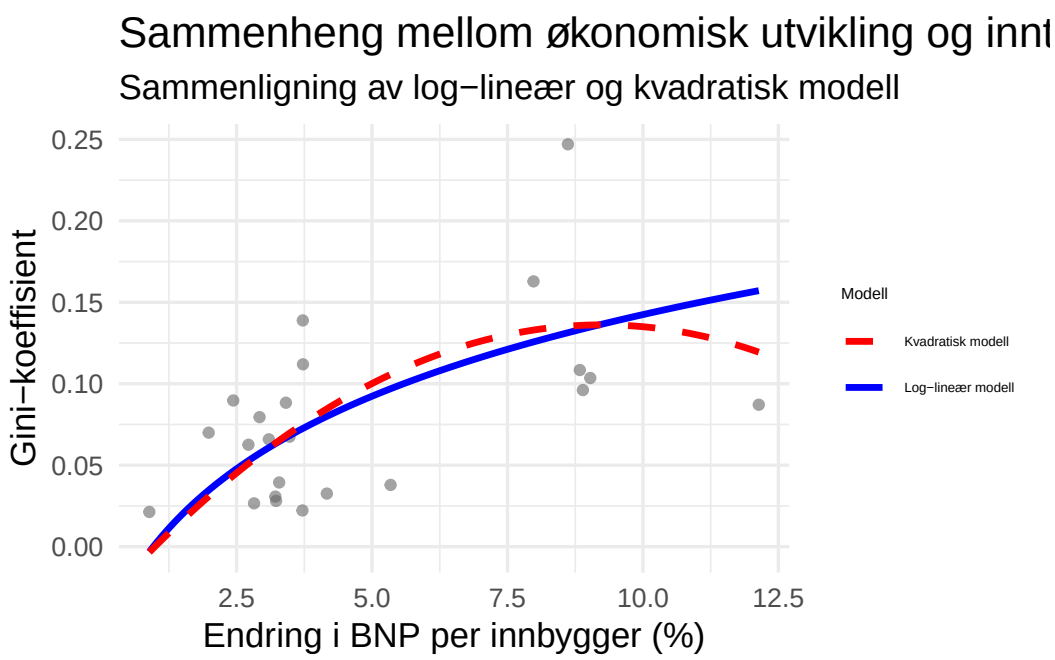

Figure 1: Visualisering av log-lineær og kvadratisk modell for sammenhengen mellom økonomisk utvikling og inntektsulikhet

Figure 1 viser sammenhengen mellom økonomisk utvikling og inntektsulikhet, der x-aksen viser endringen i BNP per innbygger (%) og y-aksen viser Gini-koeffisienten. Den log-lineære modellen (blå linje) viser en jevn og gradvis økning i ulikhet når BNP-veksten stiger. Den kvadratiske modellen (rød stiplet linje) har en tydelig buet form der ulikheten først øker og deretter flater ut eller synker svakt ved høyere vekstnivåer.

Tolkning av resultater (opp til 400 ord)

Resultatene fra Table 2 viser at den log-lineære og den kvadratiske modellen er mer overbevisende enn den kubiske modellen. Den log-lineære modellen har den laveste AIC-verdien (–76.77), mens den kvadratiske modellen har den høyeste justerte R^2 (0.528). I den log-lineære modellen er koeffisienten til log(change_GDPC_pct + 1) positiv og signifikant (Estimate = 0.0827, p = 0.030). Dette innebærer at inntektsulikheten øker når BNP-veksten stiger, men

at effekten avtar ved høyere vekstnivåer. Den kvadratiske modellen viser også en ikke-lineær sammenheng. Førsteordensleddet er positivt og signifikant (Estimate = 0.03697, p = 0.049), mens andreordensleddet er negativt (Estimate = –0.0019996, p = 0.140). Selv om det sistnevnte ikke er signifikant, antyder retningen at økningen i ulikhet avtar ved høyere vekstnivåer. Dette kan tolkes som en svak Kuznets-lignende kurve Kuznets (2019), der ulikheten først øker og senere flater ut. Den kubiske modellen gir derimot ingen forbedring. R^2 og AIC er svakere enn i de to andre modellene, og modellen tilfører dermed ikke ytterligere innsikt. Samlet sett peker resultatene mot en ikke-lineære utvikling der økonomisk vekst øker inntektsulikheten, men hvor effekten gradvis svekkes. Den log-lineære og den kvadratiske modellen fanger dette mønsteret best.

Del C: Heteroskedastisitetstesting og kausalitetsdiskusjon

Heteroskedastisitetstesting

```
# Breusch-Pagan-test
bp_log  <- bptest(model_log)
bp_kvad <- bptest(model_kvad)
bp_kub  <- bptest(model_kub)

# White-test (BP på predikerte verdier og deres kvadrat)
wt_log  <- bptest(model_log,  ~ fitted(model_log)  + I(fitted(model_log)^2))
wt_kvad <- bptest(model_kvad, ~ fitted(model_kvad) + I(fitted(model_kvad)^2))
wt_kub  <- bptest(model_kub,  ~ fitted(model_kub)  + I(fitted(model_kub)^2))

# Score-test (car::ncvTest)
sc_log  <- ncvTest(model_log)
sc_kvad <- ncvTest(model_kvad)
sc_kub  <- ncvTest(model_kub)
```

Tolk og diskuter resultatene og implikasjonene (opptil 200 ord).

Resultatene fra Table 3 viser at Breusch–Pagan-, White- og Score-testene ikke finner noen tegn til heteroskedastisitet i noen av modellene. For den log-lineære modellen er p-verdiene 0.456 (BP), 0.185 (White) og 0.205 (Score). Den kvadratiske modellen har tilsvarende høye p-verdier (0.467, 0.189 og 0.282), og selv for den kubiske modellen ligger alle p-verdiene godt over 0.05 (0.450, 0.096 og 0.166). Siden alle testene gir p-verdier over 0.05, kan ikke nullhypotesen om homoskedastisitet forkastes. Dette betyr at variansen i residualene er stabil, og at modellen ikke lider av heteroskedastisitet. Det er derfor ikke behov for å benytte robuste standardfeil, ettersom heteroskedastisitet ikke utgjør et problem i disse modellene.

Table 3: Resultater fra Breusch–Pagan-, White- og Score-testene for de tre modellene

```
# Samler testresultatene i en tabell
hetero_test <- tibble(
  Modell  = c("Log-lineær", "Kvadratisk", "Kubisk"),
  BP_p    = c(bp_log$p.value,  bp_kvad$p.value,  bp_kub$p.value),
  White_p = c(wt_log$p.value,  wt_kvad$p.value,  wt_kub$p.value),
  Score_p = c(sc_log$p,        sc_kvad$p,        sc_kub$p)
)
hetero_test
```

```
# A tibble: 3 x 4
  Modell      BP_p White_p Score_p
  <chr>      <dbl>   <dbl>   <dbl>
1 Log-lineær 0.456  0.185    0.205
2 Kvadratisk 0.467  0.189    0.282
3 Kubisk     0.450  0.0957   0.166
```

Kausalitetsdiskusjon (200–400 ord)

Selv om denne studien viser en statistisk sammenheng mellom økonomisk utvikling og inntektsulikhet, der utdanning, transport og demografi inngår som forklaringsvariabler, kan resultatene ikke tolkes som kausale. Det finnes flere grunner til dette. For det første kan det være omvendt kausalitet. Økonomisk vekst kan påvirke ulikhet, men ulikhet kan også påvirke vekst. For eksempel kan regioner med høy ulikhet ha lavere investeringer i utdanning, dårligere transportforhold og en skjev befolkningssammensetning, noe som kan redusere økonomisk utvikling. I dette tilfellet viser modellen en toveis sammenheng snarere enn en klar årsakssammenheng. For det andre kan modellen mangle viktige variabler. Selv om analysen inkluderer utdanning, veitetthet og andelen eldre, noe som gir modellen en viss forklaringskraft, mangler det fortsatt flere andre sentrale faktorer som blant annet institusjonskvalitet, arbeidsmarkedsforhold eller fordeling av kapital. Dersom slike faktorer påvirker både økonomisk vekst og ulikhet, kan resultatene bli skjeve på grunn av utelatte variabler. Et annet problem er samtidig påvirkning. Økonomisk vekst og inntektsulikhet kan påvirke hverandre i samme periode, og ikke bare i én retning. Når begge variablene endrer seg samtidig og påvirker hverandre, blir det vanskelig å skille hva som er årsak og hva som er virkning. For å identifisere kausale sammenhenger bedre, trengs mer presise analyser som kan kontrollere for regionale forhold som ikke endrer seg over tid, og dermed redusere problemet med utelatte variabler. Det er også mulig å bruke instrumentvariabler eller tidsforsinkede variabler for å redusere omvendt kausalitet og samtidig påvirkning. Flere kontrollvariabler eller lengre tidsserier kan i tillegg gjøre kausalanalysen mer pålitelig. Oppsummert viser analysen bare at det finnes en sammenheng mellom variablene, men ikke hva som er årsaken. For å si noe mer sikkert om årsakssammenhenger, trengs det mer grundige analyser som kan håndtere at viktige forhold kan mangle i modellen, eller at variablene kan påvirke hverandre.

Del D: Panel

Panelestimering

Datainnsamling for Panelestimering

```
# Laster ned og filtrerer datasett for perioden 2008-2022
# Nedlasting av BNP 2008-2022
gdp_panel <- get_eurostat("nama_10r_3gdp", time_format="num") %>%
  filter(
    unit == "MIO_EUR",
    TIME_PERIOD %in% 2008:2022,
    nchar(geo) == 5,
    substr(geo,1,2) %in% c("BE","NL","BG","NO")
  ) %>%
  select(geo, TIME_PERIOD, values_gdp = values)
```

```
indexed 0B in  0s, 0B/s
indexed 1.00TB in  0s, 57.45TB/s
```

```
# Nedlasting av befolkning 2008-2022
pop_panel <- get_eurostat("demo_r_pjanaggr3", time_format="num") %>%
  filter(
    sex == "T",
    age == "TOTAL",
    TIME_PERIOD %in% 2008:2022,
    nchar(geo) == 5,
    substr(geo,1,2) %in% c("BE","NL","BG","NO")
  ) %>%
  select(geo, TIME_PERIOD, values_pop = values)
```

```
indexed 0B in  0s, 0B/s
indexed 1.00TB in  0s, 22.57TB/s
```

```
# Slå sammen
gdp_pop_panel <- gdp_panel %>%
  left_join(pop_panel, by = c("geo", "TIME_PERIOD")) %>%
  mutate(
    NUTS2 = substr(geo, 1, 4),
    GDPC  = values_gdp * 1e6 / values_pop
  ) %>%
```

```
  filter(values_pop > 0, is.finite(GDPC)) %>%
  # beregnes årlig vekstrate i GDPC på NUTS2-nivå
  group_by(NUTS2) %>%
  arrange(TIME_PERIOD, .by_group = TRUE) %>%
  mutate(change_GDPC_pct = (GDPC / lag(GDPC) - 1) * 100) %>%
  ungroup() %>%
  # beregnes Gini på (NUTS2, TIME_PERIOD)-nivå, og oppsummeres direkte
  group_by(NUTS2, TIME_PERIOD) %>%
  mutate(
    Gini_weighted = if (n() >= 2)
      as.numeric(gini.wtd(GDPC, weights = values_pop))
    else
      NA_real_
  ) %>%
  summarise(
    GDPC           = sum(values_gdp * 1e6) / sum(values_pop),  # Befolkningsvektet GDPC for
    change_GDPC_pct = mean(change_GDPC_pct, na.rm = TRUE),      # Skal være identisk innen s
    Gini_weighted   = mean(Gini_weighted, na.rm = TRUE),        # Beholder NA, filtrerer ikk
    .groups = "drop"
  )
```

```
# høy utdanning 2008-2022
edu_panel <- utdanning %>%
  filter(
    TIME_PERIOD %in% 2008:2022,
    sex=="T", age=="Y25-64", unit=="PC",
    isced11=="ED5-8",
    nchar(geo)==4,
    substr(geo,1,2) %in% c("BE","NL","BG","NO")
  ) %>%
  select(NUTS2 = geo, TIME_PERIOD, Høy_utdanning = values)
```

```
# population (andel over 65 år) 2008-2022
pop_panel <- get_eurostat("demo_r_pjangrp3", time_format="num") %>%
  filter(
    TIME_PERIOD %in% 2008:2022,
    sex=="T",
    nchar(geo)==4,
    substr(geo,1,2) %in% c("BE","NL","BG","NO")
  ) %>%
  group_by(geo, TIME_PERIOD) %>%
  summarise(
    pop_over65 = sum(values[age %in% c("Y65-69","Y70-74","Y75-79","Y80-84","Y85-89","Y_GE85"
    pop_total  = sum(values[age=="TOTAL"]),
    andelen_over65 = pop_over65/pop_total * 100,
    .groups = "drop"
```

```
  ) %>%
  rename(NUTS2=geo)%>%
  select(NUTS2, TIME_PERIOD, andelen_over65)
```

```
indexed 0B in  0s, 0B/s
indexed 1.00TB in  0s, 18.17TB/s
```

```
# Kobler sammen datasett
panel_df <- gdp_pop_panel %>%
  left_join(edu_panel, by = c("NUTS2", "TIME_PERIOD")) %>%
  left_join(pop_panel, by = c("NUTS2", "TIME_PERIOD"))
```

Panelmodeller (Fire FE-spesifikasjoner)

```
# Oppretter landkode fra NUTS2 og aggregerer data til land × år
panel_country <- panel_df %>%
  mutate(country = substr(NUTS2, 1, 2)) %>%
  group_by(country, TIME_PERIOD) %>%
  summarise(
    Gini_weighted   = mean(Gini_weighted,   na.rm = TRUE),
    change_GDPC_pct = mean(change_GDPC_pct, na.rm = TRUE),
    Høy_utdanning   = mean(Høy_utdanning,   na.rm = TRUE),
    andelen_over65  = mean(andelen_over65,  na.rm = TRUE),
    .groups = "drop"
  )

panel_country <- pdata.frame(panel_country, index = c("country", "TIME_PERIOD"))

# Modell 1: land-faste effekter
fe_country <- plm(
  Gini_weighted ~ change_GDPC_pct + Høy_utdanning + andelen_over65,
  data   = panel_country,
  model  = "within",
  effect = "individual"
)
summary(fe_country)
```

```
Oneway (individual) effect Within Model

Call:
plm(formula = Gini_weighted ~ change_GDPC_pct + Høy_utdanning +
    andelen_over65, data = panel_country, effect = "individual",
```

```
    model = "within")

Unbalanced Panel: n = 4, T = 8-9, N = 35

Residuals:
       Min.     1st Qu.      Median     3rd Qu.        Max.
-1.8538e-02 -1.6705e-03  3.8856e-05  2.8335e-03  2.0941e-02

Coefficients:
                  Estimate  Std. Error t-value Pr(>|t|)
Høy_utdanning   0.00041389  0.00083616  0.4950   0.6243
andelen_over65 -0.00054616  0.00155307 -0.3517   0.7276

Total Sum of Squares:    0.0014812
Residual Sum of Squares: 0.0014685
R-Squared:      0.0085928
Adj. R-Squared: -0.16234
F-statistic: 0.125676 on 2 and 29 DF, p-value: 0.88238
```

```r
# Konverterer datasettet til et panelobjekt med NUTS2 som enhets-ID og TIME_PERIOD som tidsv
panel_df <- pdata.frame(panel_df, index = c("NUTS2", "TIME_PERIOD"))
```

```r
# Modell 2: Faste effekter for år (year FE)
fe_year <- plm(
  Gini_weighted ~ change_GDPC_pct + `Høy_utdanning` + andelen_over65,
  data   = panel_df,   # pdata.frame med NUTS2 og år
  model  = "within",
  effect = "time"      # år som tids-FE
)
summary(fe_year)
```

```
Oneway (time) effect Within Model

Call:
plm(formula = Gini_weighted ~ change_GDPC_pct + Høy_utdanning +
    andelen_over65, data = panel_df, effect = "time", model = "within")

Unbalanced Panel: n = 23, T = 6-9, N = 204

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-0.0801918 -0.0346544 -0.0074101  0.0193569  0.1905883

Coefficients:
                 Estimate Std. Error t-value Pr(>|t|)
Høy_utdanning  0.00010744 0.00061305  0.1753  0.86106
```

```
andelen_over65 0.00315474 0.00184516  1.7097  0.08892 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    0.54816
Residual Sum of Squares: 0.53593
R-Squared:      0.022307
Adj. R-Squared: -0.02835
F-statistic: 2.20176 on 2 and 193 DF, p-value: 0.11338
```

```r
# Modell 3: Faste effekter på NUTS2-nivå (region FE)
fe_region <- plm(
  Gini_weighted ~ change_GDPC_pct + `Høy_utdanning` + andelen_over65,
  data   = panel_df,
  model  = "within",
  effect = "individual"    # individual = NUTS2-FE siden panelet er indeksert etter NUTS2
)
summary(fe_region)
```

```
Oneway (individual) effect Within Model

Call:
plm(formula = Gini_weighted ~ change_GDPC_pct + Høy_utdanning +
    andelen_over65, data = panel_df, effect = "individual", model = "within")

Unbalanced Panel: n = 23, T = 6-9, N = 204

Residuals:
       Min.     1st Qu.      Median     3rd Qu.        Max.
-0.03735356 -0.00460678 -0.00032325  0.00416850  0.08189310

Coefficients:
                  Estimate  Std. Error t-value Pr(>|t|)
Høy_utdanning  -0.00036387  0.00044068 -0.8257   0.4101
andelen_over65  0.00047526  0.00088044  0.5398   0.5900

Total Sum of Squares:    0.026386
Residual Sum of Squares: 0.026285
R-Squared:      0.0038523
Adj. R-Squared: -0.12971
F-statistic: 0.346112 on 2 and 179 DF, p-value: 0.70791
```

```r
# Modell 4: Toveis faste effekter (NUTS2 + år)
fe_twoway <- plm(
  Gini_weighted ~ change_GDPC_pct + `Høy_utdanning` + andelen_over65,
  data   = panel_df,
```

```
  model  = "within",
  effect = "twoways"  # kontrollerer både for NUTS2- og årsfaste effekter
)
summary(fe_twoway)
```

```
Twoways effects Within Model

Call:
plm(formula = Gini_weighted ~ change_GDPC_pct + Høy_utdanning +
    andelen_over65, data = panel_df, effect = "twoways", model = "within")

Unbalanced Panel: n = 23, T = 6-9, N = 204

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-0.0324485 -0.0057489  0.0002411  0.0048400  0.0749845

Coefficients:
                  Estimate  Std. Error t-value Pr(>|t|)
Høy_utdanning  -0.00034785  0.00053200 -0.6539   0.5141
andelen_over65  0.00095596  0.00125071  0.7643   0.4457

Total Sum of Squares:    0.024253
Residual Sum of Squares: 0.024113
R-Squared:      0.0057467
Adj. R-Squared: -0.18031
F-statistic: 0.494183 on 2 and 171 DF, p-value: 0.61094
```

```
modelsummary(
  list(
    "Region FE"   = fe_region,
    "Year FE"     = fe_year,
    "Two-way FE" = fe_twoway,
    "Country FE" = fe_country
  ),
  stars = TRUE,
  fmt = 5,
  output = "markdown"
)
```

Table 4: Sammenligning av panelmodeller

	Region FE	Year FE	Two-way FE	Country FE
Høy_utdanning	-0.00036	0.00011	-0.00035	0.00041
	(0.00044)	(0.00061)	(0.00053)	(0.00084)

Table 4: Sammenligning av panelmodeller

	Region FE	Year FE	Two-way FE	Country FE
andelen_over65	0.00048	0.00315+	0.00096	-0.00055
	(0.00088)	(0.00185)	(0.00125)	(0.00155)
Num.Obs.	204	204	204	35
R2	0.004	0.022	0.006	0.009
R2 Adj.	-0.130	-0.028	-0.180	-0.162
AIC	-1242.3	-627.2	-1259.9	-247.4
BIC	-1232.3	-617.3	-1249.9	-242.8
RMSE	0.01	0.05	0.01	0.01

- $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Panelestimeringsanalyse: (opptil 400 ord)

I panelmodellen inkluderte jeg ikke transportinfrastruktur (veitetthet). Selv om variabelen endrer seg over tid, er endringen så liten at den i praksis er tidsinvariant. En faste-effekt-modell estimerer kun basert på tidsvariasjon innen samme region, og tidsinvariante variabler blir fullstendig absorbert av regionsfaste effekter. Dermed kan ikke veitetthet estimeres på en korrekt måte og er utelatt fra panelanalysen. I stedet inkluderes kun høyere utdanning og andelen av befolkningen over 65 år. Som vist i Table 4 har alle fire modellene svært svak forklaringskraft, med R^2-verdier på henholdsvis 0.004, 0.022, 0.006 og 0.009, og de justerte R^2-verdiene er negative. Dette indikerer at modellene i liten grad klarer å forklare utviklingen i regional inntektsulikhet over tid.
Videre er koeffisientene for de inkluderte variablene svært små og gjennomgående ikke-signifikante. For utdanningsnivået er p-verdiene i alle modellene klart over 0.1, blant annet p 0.41 (region FE), 0.86 (year FE), 0.51 (two-way FE) og 0.62 (country FE). Andelen eldre har også p-verdier over 0.1 i de fleste modellene, som p 0.59 (region FE)), 0.45 (two-way FE) og 0.78 (country FE), med unntak av year FE-modellen der variabelen er svakt signifikant (p 0.0889). Dette viser at disse variablene endrer seg svært lite over tid innen hver region, og derfor ikke klarer å forklare variasjonen i Gini-koeffisienten. Samlet sett viser panelresultatene at de tidsmessige endringene innen regionene er begrenset. Både Gini, utdanningsnivå og aldersstruktur er forholdsvis stabile, noe som gjør det utfordrende for modellen å identifisere tydelige tidseffekter. Dette samsvarer med de lave R^2-verdiene og manglende signifikans i Table 4, og tyder på at regionale ulikheter i større grad skyldes langsiktige strukturelle forskjeller enn kortsiktige endringer.

Panelestimeringsdiskusjon: (opptil 400 ord)

Basert på de samlede panelestimatene fremstår modellens forklaringskraft som svak. Dette henger i stor grad sammen med den begrensede tidsvariasjonen i variablene samt egenskapene ved FE modeller, som benytter endringer over tid innen samme region til å identifisere effekter. Transportinfrastruktur (veitetthet) er også en relevant forklaringsfaktor, men ble ikke inkludert i panelmodellen på grunn av modellens metodiske begrensninger. Veitetthet er i praksis en tidsinvariant variabel i analyseperioden og vil derfor bli fullstendig absorbert av

regionsfaste effekter. Dermed kan den ikke estimeres i en faste-effekt-modell. Dette innebærer at en potensielt viktig forklaringsvariabel faller bort, noe som ytterligere begrenser modellens totale forklaringskraft. Table 4 viser tydelig at modellens forklaringskraft er svært svak. I en modell med faste effekter er identifikasjonen basert på endringer over tid innenfor samme region. I dette datasettet endrer imidlertid Gini, utdanningsnivå og andelen eldre over 65 år seg svært lite i perioden 2008–2022. Fordi variablene nesten ikke varierer, har modellen svært begrenset tidsvariasjon å utnytte. Derfor blir koeffisientene meget små, for eksempel er den laveste koeffisienten for utdanning –0.00036 og for andelen eldre –0.00055, mens den laveste R^2-verdien er helt nede på 0.004. Dette viser at modellen i liten grad klarer å forklare endringer over tid. I toveis faste effekter (Two-way FE) blir mønsteret enda tydeligere. Denne modellen kontrollerer både for regionspesifikke forskjeller og årsspesifikke sjokk, noe som gjør at enda mindre tidsvariasjon står igjen til å forklare endringer i ulikhet. I toveis faste effekter (Two-way FE) er p-verdiene høye: 0.5141 for utdanningsnivå og 0.4457 for andelen eldre. Begge ligger langt over 0.1, noe som viser at disse variablene ikke har noen tydelig statistisk effekt innen regionene over tid. Samlet sett viser dette at regionale ulikheter i større grad skyldes stabile og langsiktige strukturelle forhold, og at kortsiktige endringer i utdanning, demografi og økonomisk utvikling spiller en svært begrenset rolle for utviklingen i Gini-koeffisienten i denne perioden.

Del E: Dokumenter din bruk av KI

I denne oppgaven brukte jeg kunstig intelligens som støtte i skrive- og analysearbeidet. Programvaren som ble benyttet var ChatGPT fra OpenAI (modell GPT-5.1). KI-verktøyet ble hovedsakelig brukt til å korrigere kodefeil, oversette språk, oppsummere relevant litteratur, forklare datainnhold, forstå økonometriske begreper, kontrollere om teksten var klart formulert, samt bistå med språklig forbedring slik at fremstillingen ble presis og tydelig. Gjennom prosessen hadde jeg en målrettet spørrestrategi, hvor jeg stilte flere presise spørsmål om konkrete temaer, for eksempel betydningen av faste effekter, toveis faste effekter, hvorfor tidsinvariante variabler ikke kan estimeres i FE-modeller, og hvordan panelresultater bør tolkes. Jeg brukte også KI til å oversette enkelte avsnitt mellom kinesisk og norsk for å sikre at språket var konsistent i hele rapporten. Etter hver redigering kontrollerte igjen jeg innholdet selv for å unngå overredigering og for å sikre at all analyse og alle konklusjoner var basert på min egen forståelse og mine egne data.

Kuznets, Simon. 2019. “Economic Growth and Income Inequality.” In *The Gap Between Rich and Poor*, 25–37. Routledge.

Lessmann, Christian, and André Seidel. 2017. “Regional Inequality, Convergence, and Its Determinants – a View from Outer Space.” *European Economic Review* 92: 110–32. https://doi.org/https://doi.org/10.1016/j.euroecorev.2016.11.009.