Assignment 2, MSB104-gruppe 6 2025

Karin Liang

Introduksjon

Denne oppgaven undersøker fortsatt sammenhengen mellom regional utvikling og ulikhet i de fire landene Belgia (BE), Nederland (NL), Bulgaria (BG) og Norge (NO). Analysen bygger delvis videre på datasettene gdp_pop og gini fra Assignment 1, men inkluderer også nye variabler og data for året 2017.

```
# Parkker
library(dplyr)
library(eurostat)
library(ggplot2)
library(readxl)
library(broom)
library(lmtest)   # Inneholder dwtest(), bptest()
library(car)      # oen diagnostiske verktøy
library(ggfortify)

# Følgende datasettene gdp_pop og gini fra assignment 1
# leser excel gdp_pop og gini
gdp_pop <- read_excel("../gdp_pop.xlsx", sheet = "Sheet 1")
gini <- read_excel("../gini.xlsx", sheet = "Sheet 1")
```

Del A

Laster ned og slår sammen datasettene for 2017

```
#  henter bnp og population for 2016 og 2017
gdp_pop <- gdp_pop %>%
  mutate(NUTS2 = substr(geo, 1, 4)) %>%
  filter(
    TIME_PERIOD %in% c(2016, 2017),
    unit_gdp == "MIO_EUR",
    unit_pop == "NR",
    sex == "T", age == "TOTAL"
  ) %>%
  group_by(NUTS2, TIME_PERIOD) %>%
  summarise(
    values_gdp = sum(values_gdp, na.rm = TRUE),   # Beholder opprinnelig
kolonnenavn
    values_pop = sum(values_pop, na.rm = TRUE),   # Beholder opprinnelig
kolonnenavn
    .groups = "drop"
  ) %>% # Summerer BNP og befolkning for hver NUTS2-region og år
```

```
  mutate(GDPC = (values_gdp * 1e6) / values_pop)   # BNP per innbygger (i
euro)
gdp_pop
```

```
# A tibble: 66 × 5
   NUTS2 TIME_PERIOD values_gdp values_pop   GDPC
   <chr>       <dbl>      <dbl>      <dbl>  <dbl>
 1 BE10         2016     80092.    1201285 66672.
 2 BE10         2017     82258.    1199095 68600.
 3 BE21         2016     82295.    1828927 44996.
 4 BE21         2017     85078.    1838863 46266.
 5 BE22         2016     26782.     866970 30891.
 6 BE22         2017     27865.     869664 32042.
 7 BE23         2016     52027.    1489084 34939.
 8 BE23         2017     54304.    1498483 36240.
 9 BE24         2016     46550.    1122600 41467.
10 BE24         2017     48159.    1130644 42594.
# ℹ 56 more rows
```

```
# beregner endringsprosenten
gdp_pop_change_2017 <- gdp_pop %>%
  arrange(NUTS2, TIME_PERIOD) %>%        # Sorterer etter region og år
  group_by(NUTS2) %>%                    # Beregner endring per region
  mutate(
    change_GDPC_pct = (GDPC / lag(GDPC) - 1) * 100   # Prosentvis endring
(%)
  ) %>%
  filter(TIME_PERIOD == 2017) %>%        # Beholder kun endringen for 2017
  ungroup()
gdp_pop_change_2017
```

```
# A tibble: 33 × 6
   NUTS2 TIME_PERIOD values_gdp values_pop   GDPC change_GDPC_pct
   <chr>       <dbl>      <dbl>      <dbl>  <dbl>           <dbl>
 1 BE10         2017     82258.    1199095 68600.            2.89
 2 BE21         2017     85078.    1838863 46266.            2.82
 3 BE22         2017     27865.     869664 32042.            3.73
 4 BE23         2017     54304.    1498483 36240.            3.72
 5 BE24         2017     48159.    1130644 42594.            2.72
 6 BE25         2017     45242.    1188407 38069.            3.10
 7 BE31         2017     18176.     399735 45470.            4.37
 8 BE32         2017     33010.    1342053 24596.            2.92
 9 BE33         2017     30800.    1106039 27847.            3.48
10 BE34         2017      6864.     284617 24118.            3.41
# ℹ 23 more rows
```

```
# Velger Gini-vektene for 2017
gini_2017 <- gini %>%
```

```
  mutate(TIME_PERIOD = as.integer(TIME_PERIOD)) %>%  # Konverterer kolonnen
TIME_PERIOD til heltallstype
  filter(TIME_PERIOD == 2017)
gini_2017
```

```
# A tibble: 33 × 5
   country NUTS2 TIME_PERIOD n_nuts3 Gini_weighted
   <chr>   <chr>       <int>   <dbl>         <dbl>
 1 BE      BE10         2017       1       NA
 2 BE      BE21         2017       3        0.0266
 3 BE      BE22         2017       3        0.112
 4 BE      BE23         2017       6        0.139
 5 BE      BE24         2017       2        0.0626
 6 BE      BE25         2017       8        0.0659
 7 BE      BE31         2017       1       NA
 8 BE      BE32         2017       7        0.0795
 9 BE      BE33         2017       5        0.0675
10 BE      BE34         2017       5        0.0883
# ℹ 23 more rows
```

```
# Slå sammen datasettet og kall det data_2017
data_2017_a <- gdp_pop_change_2017 %>%
  left_join(gini_2017, by = "NUTS2") %>%
  filter(!is.na(Gini_weighted))
data_2017_a
```

```
# A tibble: 23 × 10
   NUTS2 TIME_PERIOD.x values_gdp values_pop   GDPC change_GDPC_pct country
   <chr>         <dbl>      <dbl>      <dbl>  <dbl>           <dbl> <chr>
 1 BE21           2017     85078.    1838863 46266.            2.82 BE
 2 BE22           2017     27865.     869664 32042.            3.73 BE
 3 BE23           2017     54304.    1498483 36240.            3.72 BE
 4 BE24           2017     48159.    1130644 42594.            2.72 BE
 5 BE25           2017     45242.    1188407 38069.            3.10 BE
 6 BE32           2017     33010.    1342053 24596.            2.92 BE
 7 BE33           2017     30800.    1106039 27847.            3.48 BE
 8 BE34           2017      6864.     284617 24118.            3.41 BE
 9 BE35           2017     13103.     494127 26517.            2.44 BE
10 BG31           2017      3540.     753059  4701.           12.1  BG
# ℹ 13 more rows
# ℹ 3 more variables: TIME_PERIOD.y <int>, n_nuts3 <dbl>, Gini_weighted
<dbl>
```

Først hentes data for BNP og population for årene 2016 og 2017 fra datasettet gdp_pop. Deretter beregnes endringen og den prosentvise endringen av BNP per innbygger i hver NUTS2-region fra 2016 til 2017.
Gini-koeffisientene for 2017 hentes fra datasettet gini, og disse kobles sammen med BNP-dataene ved hjelp av NUTS2-koden som felles nøkkel. Resultatet er et kombinert datasett data_2017_a som brukes i den videre enkle lineære regresjonsanalysen

Enkel lineær regresjonsmodell

Her brukes den avhengige variabelen Gini_weighted og den uavhengige variabelen change_GDPC_pct for å estimere en enkel lineær regresjonsmodell.

```
# Avhengig variabel: Gini_weighted
# Uavhengig variabel: change_GDPC_pct
# Enkel lineær regresjonsmodell
model_simple <- lm(Gini_weighted ~ change_GDPC_pct, data = data_2017_a)

# Sjekker dataene direkte
summary(model_simple)
```

```

Call:
lm(formula = Gini_weighted ~ change_GDPC_pct, data = data_2017_a)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06377 -0.03334 -0.01026  0.02047  0.13044

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.032586   0.018641   1.748  0.09505 .
change_GDPC_pct 0.009746   0.003353   2.906  0.00844 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04598 on 21 degrees of freedom
Multiple R-squared:  0.2869,    Adjusted R-squared:  0.2529
F-statistic: 8.447 on 1 and 21 DF,  p-value: 0.008441
```

Table 1: Resultater fra enkel lineær regresjon for 2017 (NUTS2)

```
tidy(model_simple)
```

```
# A tibble: 2 × 5
  term            estimate std.error statistic p.value
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)      0.0326    0.0186       1.75 0.0951
2 change_GDPC_pct  0.00975   0.00335      2.91 0.00844
```

Table 2: Modelltilpasningsstatistikk for regresjonen i 2017

```
glance(model_simple)
```

```
# A tibble: 1 × 12
  r.squared adj.r.squared  sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl>  <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.287         0.253 0.0460      8.45 0.00844     1   39.2 -72.5 -69.1
# ℹ 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Modellspesifikasjon

Den enkle lineære regresjonsmodellen analyserer hvordan regional økonomisk utvikling påvirker inntektsulikhet på NUTS2-nivå i 2017. Den avhengige variabelen er Gini_weighted, som måler inntektsulikheten, jo høyere Gini-koeffisienten er, desto større er graden av ulikhet. Den uavhengige variabelen er den prosentvise endringen i BNP per innbygger fra 2016 til 2017, som representerer regional økonomisk utvikling. Dette tilsvarer en av syv dimensjonener utvikling i modellen til C. Lessmann and A. Seidel [1]. Variabelen måler økonomisk vekst og brukes som en indikator på regional utvikling, som igjen kan påvirke ulikhet avhengig av regionens utviklingsnivå.

C. Lessmann and A. Seidel [1] peker på at forholdet mellom økonomisk utvikling og ulikhet følger et N-formet mønster, der ulikheten øker i lavinntektsland, reduseres i mellominntektsland og stiger svakt igjen i høyinntektsland. Dette innebærer at økonomisk vekst ikke automatisk fører til mindre ulikhet. Ulikhet bestemmes ikke bare av økonomisk vekst, men påvirkes også av flere strukturelle faktorer. Ifølge C. Lessmann and A. Seidel [1] avhenger graden av ulikhet av syv hoveddimensjoner: utvikling, mobilitet, åpenhet, ressurser, institusjoner, overføringer og utdanning, og etnisitet. Disse faktorene virker sammen og kan forklare hvorfor regionale forskjeller i ulikhet oppstår, utover effekten av selve den økonomiske veksten.

I tråd med teorien om en N-formet i sammenheng forventes koeffisienten for prosentvis endring i BNP per innbygger å være svakt positiv. Dette vil si at regioner med høyere økonomisk vekst kan få noe større ulikhet dersom veksten hovedsakelig skjer i enkelte sektorer eller byområder. En svak eller ikke-signifikant sammenheng kan derimot tyde på at mange regioner allerede opplever inntektskonvergens, der forskjellene mellom regionene gradvis blir mindre.

Modelldiagnostikk

Den enkel lineær regresjonsmodellen analyserer hvordan økonomisk utvikling(endringen i BNP per innbygger) påvirker regional inntektsulikhet målt ved Gini-koeffisienten på NUTS2-nivå i 2017.

Table 2 viser en R-squared på 0.2869 og en justert R-squared på 0.2529. R-squared viser at variabelen endringen i BNP per innbygger forklarer om lag 28.7 % av ulikheten mellom regionene. En justert R-squared på 0.2529 betyr at, etter justering for antall variabler, modellen forklarer om lag 25 % av ulikheten mellom NUTS2-regionene i 2017. Selv om forklaringskraften er moderat, er den tilfredsstillende for tverrsnittsdata på regionalt nivå.

En p-verdi på ca. 0.008 vist i Table 1 indikerer at modellen som helhet er statistisk signifikant. Det betyr at den uavhengige variabelen den prosentvise endringen i BNP per innbygger (change_GDPC_pct) bidrar signifikant til å forklare variasjonen i ulikhet (Gini_weighted) mellom regionene i de fire landene.

Den esiduelle standardfeilen på 0.04598 fra summary, indikerer at avvikene mellom observerte og predikerte verdier er relativt små, noe som støtter at modellen gir en rimelig god tilpasning til dataene.

Samlet sett tyder disse resultatene på at modellen har en tilfredsstillende forklaringskraft, og at endring i BNP per innbygger spiller en signifikant rolle i å forklare variasjoner i regional ulikhet i 2017.

OLS

Her vurderes de fem hovedforutsetningene for OLS: linearitet, uavhengighet, homoskedastisitet, fravær av multikollinearitet og normalfordeling av residualer.

```
# (1) Residualplott
plot(data_2017_a$change_GDPC_pct, resid(model_simple),
     main = "Residualer vs. tilpassede verdier (sjekk for linearitet)",
     xlab = "Endring i BNP per innbygger",
     ylab = "Residualer")
abline(h = 0, col = "red", lwd = 2)
```

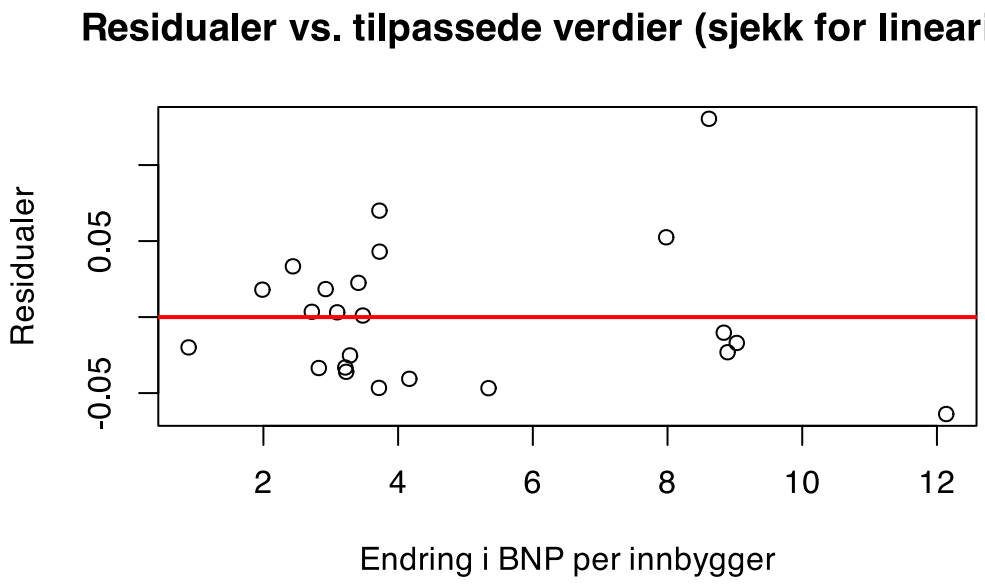

Figure 1: Residualplott for enkel lineær regresjon

Figure 1 viser en svak, men ikke systematisk kurveform. Punktene er jevnt spredt rundt nullinjen, noe som tyder på et tilnærmet lineært forhold mellom den avhengige variabelen Gini og den uavhengige variabelen endringsprosenten i BNP per innbygger. Dette støtter antakelsen om linearitet i modellen.

Table 3: Durbin–Watson-test for uavhengighet av residualer

```
# (2) Uavhengighet av residualer
# Durbin–Watson-test (nullhypotese: ingen autokorrelasjon)
dwtest(model_simple)
```

```

    Durbin-Watson test

data:  model_simple
DW = 1.3995, p-value = 0.04983
alternative hypothesis: true autocorrelation is greater than 0
```

Table 3 gir DW på 1.3995 og p-verdien på 0.0498, som indikerer en svak positiv autokorrelasjon i residualene. Dette betyr at observasjoner fra nærliggende regioner kan være delvis avhengige av hverandre.

Table 4: Breusch–Pagan-test for homoskedastisitet (konstant varians)

```
# (3) Homoskedastisitet
# Breusch–Pagan-test (sjekker for konstant varians)
bptest(model_simple)
```

```

    studentized Breusch-Pagan test

data:  model_simple
BP = 3.286, df = 1, p-value = 0.06987
```

Table 4 gir p-verdien på 0.0699, som er litt over signifikansnivået. Dette betyr at vi ikke forkaster nullhypotesen om konstant varians.Scale–Location-plottet i Figure 3 viser også en jevn spredning av punkter og en forholdsvis flat blå linje, noe som støtter antakelsen om homoskedastisitet i modellen.

```
# (4) Ingen multikollinearitet
# Denne modellen har kun en uavhengig variabel, så dette kravet er
automatisk oppfylt.
# For multippel regresjon:
# vif(model_multiple)
```

Modellen har kun en uavhengig variabel, derfor er multikollinearitet ikke et problem i denne analysen.

```
# (5) Normalfordeling av residualer
# QQ-plott + Shapiro–Wilk-test for normalitet
qqnorm(resid(model_simple))
qqline(resid(model_simple), col = "red", lwd = 2)
```

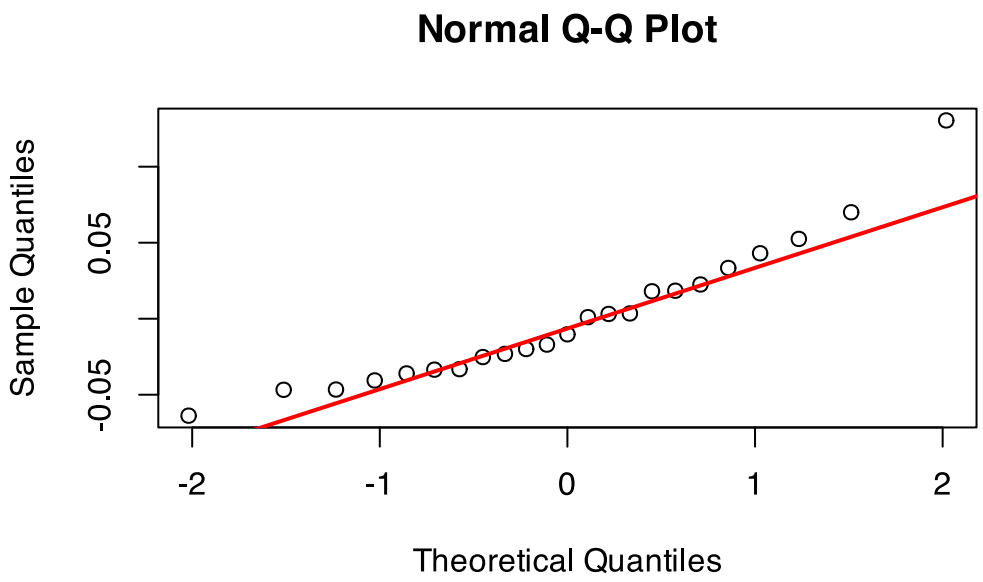

Figure 2: QQ-plott av residualene

Table 5: Shapiro–Wilk-test

```
shapiro.test(resid(model_simple))
```

```
    Shapiro-Wilk normality test

data:  resid(model_simple)
W = 0.92144, p-value = 0.07153
```

Figure 2 viser noe avvik i endene, men residualene følger for øvrig linjen godt. Table 5 gir p-verdien på 0.0715, som er større enn 0.05, og dermed kan residualene anses å være tilnærmet normalfordelte.

```
# Diagnostiske plott for OLS-modellen
autoplot(model_simple)
```

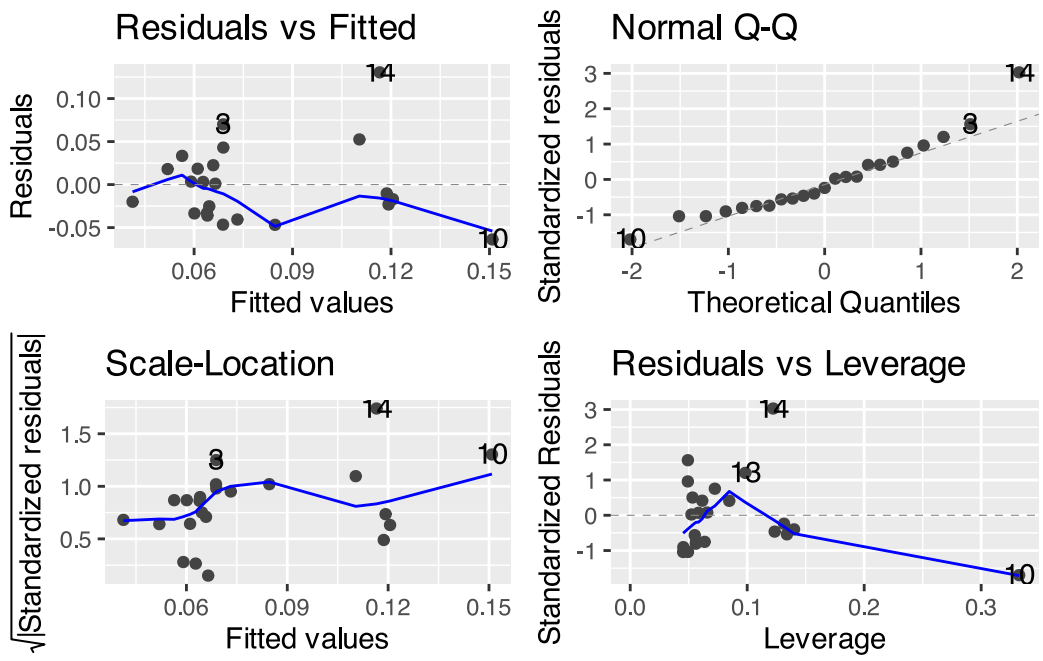

Figure 3: Diagnostiske plott for OLS-modellen

Figure 3 indikerer at observasjonene 10 og 14 har noe høy innflytelse, men ingen alvorlige avvik ble observert. Dette betyr at enkelte regioner påvirker modellen mer enn andre, men ikke i en grad som endrer hovedresultatene.

De fleste OLS-forutsetningene er tilfredsstillende oppfylt. Modellen viser et lineært forhold, ingen alvorlig heteroskedastisitet, og residualene er tilnærmet normalfordelte. Dette tyder på at modellen er pålitelig og gir gyldige estimater.

Hvis forutsetningene ikke er oppfylt, kan beregningene av standardfeil bli unøyaktige. Dette kan føre til at testen for signifikans ikke blir helt pålitelig. For å løse dette kan man bruke robuste standardfeil, eller lage en mer detaljert modell med flere forklaringsvariabler som tar hensyn til forskjeller mellom regionene.

Visuzlization

```
figur_utvikling_ulikhet <- data_2017_a %>%
  ggplot(aes(x = change_GDPC_pct, y = Gini_weighted)) +
  geom_point(color = "darkblue") +
  geom_smooth(
    method = "lm", # viser en lineær regresjonsmodell
    se = TRUE,       # et bånd rundt regresjonslinjen som representerer
standardfeilen (95 % konfidensintervall)
    color = "red",
    fill = "grey70", # skyggen grå
    linewidth = 1
  ) +
  labs(
    title = "Regional utvikling og ulikhet (2017)",
    x = "Endringprosentvis i BNP per innbygger",
    y = "Regional ulikhet (Gini-koeffisient)"
```

```
  ) +
  theme_minimal()
figur_utvikling_ulikhet
```

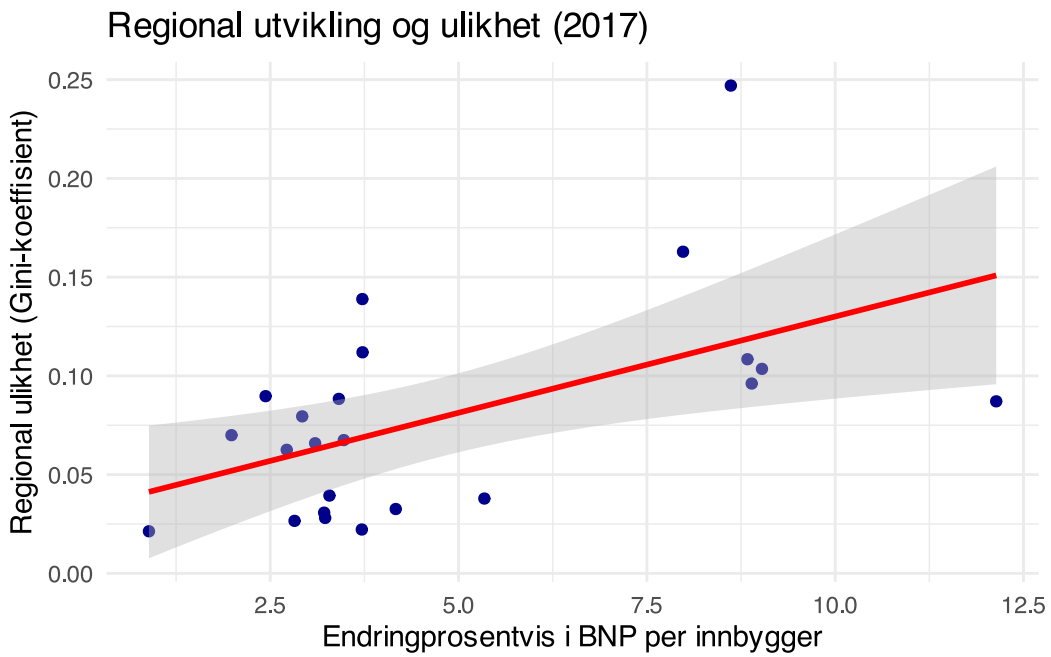

Figure 4: Forholdet mellom regional utvikling og regional ulikhet, 2017 (NUTS2)

Figure 4 viser sammenhengen mellom økonomisk vekst (prosentvis endringen i BNP per innbygger) og regional ulikhet (Gini-koeffisient) på NUTS2-nivå i 2017. Hver prikk representerer en region, mens den røde linjen viser den estimerte lineære sammenhengen, og det grå området er 95% konfidensintervallet.

Den røde linjen viser en positiv sammenheng mellom økonomisk vekst og ulikhet. Regioner med høyere vekst har som regel høyere Gini-koeffisient. Sammenhengen er moderat, men statistisk signifikant, i tråd med regresjonsresultatet der koeffisienten for change_GDPC_pct er positiv og signifikant (p = 0.0084). Dette tyder på at økonomisk vekst ikke nødvendigvis fører til mindre ulikhet, men tvert imot kan bidra til større forskjeller mellom regionene.

Del B

Datainnsamling av tre passende variabler

Her lastet jeg ned datasettene fra Eurostat for utdanning, transport og population, og valgte én variabel fra hvert datasett for videre analyse. Datasettene dekker regioner på NUTS2-nivå i Belgia, Bulgaria, Nederland og Norge.

Utdanning (høy utdanning)

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

utdanning_2017_høynivå <- utdanning %>%
```

```
  filter(
    TIME_PERIOD == 2017,
    isced11 == "ED5-8",     # Utdanningsnivå: høyere utdanning (nivå 5–8)
    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")
  ) %>%
  select(sex, unit, geo, values)
utdanning_2017_høynivå
```

```
# A tibble: 35 × 4
   sex   unit  geo   values
   <chr> <chr> <chr>  <dbl>
 1 T     PC    BE10    46.6
 2 T     PC    BE21    38.2
 3 T     PC    BE22    35.4
 4 T     PC    BE23    42.6
 5 T     PC    BE24    50.7
 6 T     PC    BE25    37.4
 7 T     PC    BE31    57.1
 8 T     PC    BE32    31.1
 9 T     PC    BE33    35.6
10 T     PC    BE34    37.1
# ℹ 25 more rows
```

```
# Kolonnen 'values' er målt i prosent (unit = "PC") og kan brukes direkte i
analysen
```

Datasettet utdanning (edat_lfse_04) er hentet fra Eurostat, og variabelen andel personer med høyere utdanning (ISCED 5–8) er valgt. Datasettet ble filtrert slik at det omfatter begge kjønn, personer i alderen 25–64 år og verdier målt i prosent av befolkningen. Deretter ble datasettet utdanning_2017_høynivå opprettet.

Transport (veitetthet)

```
# Laster ned transportdatasettet fra Eurostat
transport <- get_eurostat("tran_r_net", time_format = "num")

transport_2017 <- transport %>%
  filter(
    TIME_PERIOD == 2017,
    tra_infr == "MWAY",         # Variabel: motorveinett (motorways)
    unit == "KM",              # Enhet: kilometer (total lengde)
    nchar(geo) == 4,
    substr(geo, 1, 2) %in% c("BE","BG","NL","NO")
  ) %>%
  select(tra_infr, unit, geo, values)
transport_2017
```

```
# A tibble: 40 × 4
   tra_infr unit  geo   values
   <chr>    <chr> <chr>  <dbl>
 1 MWAY     KM    BE10      11
 2 MWAY     KM    BE21     220
 3 MWAY     KM    BE22     106
 4 MWAY     KM    BE23     196
 5 MWAY     KM    BE24     175
 6 MWAY     KM    BE25     187
 7 MWAY     KM    BE31      63
 8 MWAY     KM    BE32     284
 9 MWAY     KM    BE33     266
10 MWAY     KM    BE34     154
# ℹ 30 more rows
```

```
# Her er variabelen 'values' lengden på motorveinettet (i km).
# Laster også ned arealdata for regionene, for å sammenligne regioner, må
vi beregne tetthet (km per km²),

# Laster ned arealdata fra Eurostat
area <- get_eurostat("reg_area3", time_format = "num")
area_2017 <- area %>%
  filter(TIME_PERIOD == 2017,
         unit == "KM2",
         nchar(geo) == 4,
         substr(geo,1,2) %in% c("BE","BG","NL","NO")) %>%
  group_by(geo) %>%
  summarise(area_km2 = max(values, na.rm = TRUE), .groups = "drop") %>% #
Tar største verdi hvis duplikater finnes
  mutate(area_km2 = ifelse(is.infinite(area_km2), NA, area_km2)) %>%  #
Erstatter -Inf med NA
  filter(!is.na(area_km2))   # Fjerner manglende verdier
area_2017
```

```
# A tibble: 36 × 2
   geo   area_km2
   <chr>    <dbl>
 1 BE10       162
 2 BE21      2876
 3 BE22      2428
 4 BE23      3008
 5 BE24      2119
 6 BE25      3167
 7 BE31      1097
 8 BE32      3814
 9 BE33      3858
10 BE34      4460
# ℹ 26 more rows
```

```r
# Beregner veitetthet = motorveilengde / areal * 1000 (for å få km per 1000
km²)
transport_2017tetthet <- transport_2017 %>%
  left_join(area_2017, by = "geo") %>%
  mutate(tetthet = values / area_km2 * 1000) %>%  # Veilengde ÷ areal ×
1000
  select(geo, tetthet)%>%
  filter(is.finite(tetthet)) # Fjerner uendelige eller manglende verdier
transport_2017tetthet
```

```
# A tibble: 36 × 2
   geo   tetthet
   <chr>   <dbl>
 1 BE10     67.9
 2 BE21     76.5
 3 BE22     43.7
 4 BE23     65.2
 5 BE24     82.6
 6 BE25     59.0
 7 BE31     57.4
 8 BE32     74.5
 9 BE33     68.9
10 BE34     34.5
# ℹ 26 more rows
```

Datasettet transport (tran_r_net) er hentet fra Eurostat, og variabelen motorveinett (MWAY) ble valgt. Datasettet ble filtrert slik at det omfatter regioner på NUTS2-nivå i Belgia, Bulgaria, Nederland og Norge, og verdiene viser den totale lengden av motorveinettet (KM) i 2017. Siden denne variabelen måler absolutt lengde i kilometer, egner den seg ikke direkte til videre analyse. Derfor ble datasettet areal (reg_area3) hentet fra Eurostat, som inneholder informasjon om arealet til hver region (målt i km²). Deretter ble det filtrert slik at det kun omfatter NUTS2-regionene i Belgia, Bulgaria, Nederland og Norge, og dannet datasettet area_2017. Dette ble brukt til å beregne veitetthet (tetthet) som motorveilengde ÷ areal × 1000 (km per 1000 km²).Deretter ble datasettet transport_2017tetthet dannet, som viser veinettetets tetthet i hver region.

Population (andel av befolkningen 65 år og eldre)

```r
# Laster ned demografidatasettet fra Eurostat
population <- get_eurostat("demo_r_pjangrp3", time_format = "num")

population_2017_over65 <- population %>%
  filter(
    TIME_PERIOD == 2017,
    sex == "T",
    nchar(geo) == 4,
    substr(geo, 1, 2) %in% c("BE", "BG", "NL", "NO")
  ) %>%
  group_by(geo) %>%    # Grupperer etter region
  summarise(
```

```
    pop_over65 = sum(values[age %in%
c("Y65-69","Y70-74","Y75-79","Y80-84","Y85-89","Y_GE85")], na.rm = TRUE),
    pop_total = sum(values[age == "TOTAL"], na.rm = TRUE)
  ) %>%
  # Beregner andelen av befolkningen som 65 år og eldre (i prosent)
  mutate(andelen_over65 = pop_over65 / pop_total * 100) %>%
  ungroup() %>%
  filter(is.finite(andelen_over65)) #Fjerner manglende eller uendelige
verdier
population_2017_over65
```

```
# A tibble: 39 × 4
   geo   pop_over65 pop_total andelen_over65
   <chr>      <dbl>     <dbl>          <dbl>
 1 BE10      173045   1199095           14.4
 2 BE21      381090   1838863           20.7
 3 BE22      182124    869664           20.9
 4 BE23      318912   1498483           21.3
 5 BE24      235760   1130644           20.9
 6 BE25      296196   1188407           24.9
 7 BE31       79882    399735           20.0
 8 BE32      266344   1342053           19.8
 9 BE33      220696   1106039           20.0
10 BE34       51433    284617           18.1
# ℹ 29 more rows
```

```
# Her beregnes andelen av befolkningen som er 65 år eller eldre.
# Siden befolkningsstørrelsen varierer mye mellom regionene
# analyserer ikke antall personer direkte, men prosentandelen eldre i
befolkningen.
# Dette er en sentral strukturell faktor som kan påvirke inntektsulikhet.
```

Datasettet population (demo_r_pjangrp3) er hentet fra Eurostat og inneholder demografisk informasjon for europeiske regioner. Datasettet ble filtrert slik at det omfatter regioner på NUTS2-nivå i Belgia, Bulgaria, Nederland og Norge i 2017, og inkluderer begge kjønn. Ettersom befolkningsstørrelsen varierer betydelig mellom regionene, analyseres ikke antall personer direkte, men prosentandelen eldre i befolkningen. Variabelen måler andelen av befolkningen som er 65 år og eldre, beregnet som antall personer over 65 år delt på totalbefolkningen i regionen, multiplisert med 100 for å få prosent. Datasettet population_2017_over65 ble dermed dannet for videre analyse.

Begrunner valgene mine (200–400 ord)

Jeg velger andelen personer med høyere utdanning som analysevariabel, fordi D. Coady and A. Dizioli [2] viser studien at den positive sammenhengen mellom utdanningsulikhet og inntektsulikhet blir betydelig sterkere, og bekrefter at utvidelse av utdanning reduserer inntektsulikhet ved å redusere ulikheten i utdanning. Dermed forventes andelen personer med høyere utdanning å ha en negativ sammenheng med Gini-koeffisienten, ettersom et høyere utdanningsnivå bidrar til å utjevne inntektsforskjeller og redusere ulikheten mellom regioner.

Dermed representerer andelen personer med høyere utdanning utdanningsdata for videre analyser, der det antas at denne variabelen kan påvirke regional ulikhet.

Jeg velger veitetthet som analysevariabel, fordi C. Calderón and L. Servén [3] viser at de to sentrale resultatene er at økonomisk vekst påvirkes positivt av beholdningen av infrastrukturaktiva, og at inntektsulikhet reduseres når mengden og kvaliteten på infrastrukturen øker. Videre viser C. Calderón and L. Servén [3] også at veinettets tetthet har en tydelig negativ sammenheng med inntektsulikhet, og peker at Gini-koeffisienten er negativt korrelert med infrastrukturnivået for veier (-0,48). Dermed forventes variabelen veitetthet å ha en negativ sammenheng med Gini-koeffisienten. Dette tyder på at bedre og tettere transportinfrastruktur kan bidra til lavere regional ulikhet. Dermed representerer veitetthet transportdata for videre analyser av hvordan jeg antar at det kan påvirke regional ulikhet.

Jeg velger andelen av befolkningen som er 65 år og eldre som analysevariabel, fordi M. Dolls, K. Doorley, A. Paulus, H. Schneider, and E. Sommer [4] viser at demografiske endringer sannsynligvis vil føre til økende ulikhet, og at en større andel eldre i befolkningen har en tendens til å øke den samlede inntektsulikheten. Dette innebærer at en økende andel eldre kan bidra til høyere inntektsforskjeller. Dermed representerer andelen personer over 65 år populationdata for videre analyser av hvordan demografiske endringer kan påvirke regional ulikhet.

Samlet sett har jeg valgt disse tre variablene, andelen personer med høyere utdanning (utdanning), veitetthet (transport) og andelen av befolkningen som er 65 år og eldre (population), for å inkluderes i en multippel lineær regresjonsmodell for å analysere hvordan disse tre faktorene kan påvirke regional ulikhet.

Multippel lineær regresjonsmodell

Slår sammen datasettene

```
# Velger relevante variabler fra hvert datasett og slår dem sammen

# Velger endring i BNP per innbygger (%) og vektet Gini-koeffisient
data_2017_a <- data_2017_a %>%
  select(NUTS2, country, change_GDPC_pct, Gini_weighted)

# Utdanningsdata (andel av befolkningen med høyere utdanning)
utdanning_2017_høynivå <- utdanning_2017_høynivå %>%
  rename(NUTS2 = geo, utdanning_hoy = values) %>%
  select(NUTS2, utdanning_hoy)

# Transportdata (veitetthet))
transport_2017tetthet <- transport_2017tetthet %>%
  rename(NUTS2 = geo, vei_tetthet = tetthet)

# Demografidata (andel av befolkningen over 65 år))
population_2017_over65 <- population_2017_over65 %>%
  rename(NUTS2 = geo) %>%
  select(NUTS2, andelen_over65)

# Slår sammen alle datasettene basert på NUTS2
```

```r
data_2017_b <- data_2017_a %>%
  left_join(utdanning_2017_høynivå, by = "NUTS2") %>%
  left_join(transport_2017tetthet, by = "NUTS2") %>%
  left_join(population_2017_over65, by = "NUTS2")
```

Her velges hovedvariabler fra hvert datasett, og alle de relevante datasettene for året 2017 slås sammen på NUTS2-nivå, slik at variablene kan analyseres i ett felles datasett. Datasettet data_2017_a beholder prosentvis endringen i BNP per innbygger og den vektede Gini-koeffisienten. Videre hentes variabelen for andelen av befolkningen med høyere utdanning fra utdanningsdataene, variabelen for veitetthet fra transportdataene, og variabelen for andelen av befolkningen som er 65 år og eldre fra populationdataene. Etter at datasettene er slått sammen, dannes datasettet data_2017_b, som inneholder alle variablene som skal brukes i multippel lineær regresjonsanalyse.

Multippel lineær regresjonsmodell

```r
# Estimerer en multippel lineær regresjonsmodell
model_multiple <- lm(
  Gini_weighted ~ change_GDPC_pct + utdanning_hoy + vei_tetthet +
andelen_over65,
  data = data_2017_b)
summary(model_multiple)
```

```
Call:
lm(formula = Gini_weighted ~ change_GDPC_pct + utdanning_hoy +
    vei_tetthet + andelen_over65, data = data_2017_b)

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 .
utdanning_hoy    0.0043338  0.0014869   2.915  0.00966 **
vei_tetthet     -0.0009082  0.0004069  -2.232  0.03934 *
andelen_over65  -0.0028665  0.0043247  -0.663  0.51634
---
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
```

Table 6: Resultater fra multippel lineær regresjonsmodell for 2017 (NUTS2)

```
tidy(model_multiple)   # Koeffisienter, standardfeil og p-verdier
```

```
# A tibble: 5 × 5
  term             estimate std.error statistic p.value
  <chr>               <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)     -0.0142    0.117       -0.121 0.905
2 change_GDPC_pct  0.0113    0.00553      2.04  0.0574
3 utdanning_hoy    0.00433   0.00149      2.91  0.00966
4 vei_tetthet     -0.000908  0.000407    -2.23  0.0393
5 andelen_over65  -0.00287   0.00432     -0.663 0.516
```

Table 7: Modelltilpasningsstatistikk for multippel lineær regresjonen i 2017

```
glance(model_multiple) # R², justert R² og statistikk
```

```
# A tibble: 1 × 12
  r.squared adj.r.squared  sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl>  <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.586         0.489 0.0378      6.02 0.00331     4   43.7 -75.3 -68.8
# ℹ 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Dokumenter og begrunner modellspesifikasjonen min (200–400 ord)

Den multippel lineære regresjonsmodellen analyserer hvordan økonomisk utvikling (endringen i BNP per innbygger), utdanning (andelen av befolkningen med høyere utdanning), transport (veitetthet) og population (andelen av befolkningen som er 65 år og eldre) påvirker regional inntektsulikhet målt ved Gini-koeffisienten på NUTS2-nivå i 2017.

Variabelen for høyere utdanning har en positiv og signifikant koeffisient, og er den mest signifikante variabelen i modellen. Dette samsvarer med D. Coady and A. Dizioli [2], som viser at et høyere utdanningsnivå bidrar til å redusere inntektsforskjeller.

Variabelen veitetthet har en negativ estimate og en signifikant koeffisient. Dette er i tråd med teorien til C. Calderón and L. Servén [3], som viser at transportinfrastruktur fremmer økonomisk integrasjon og kan bidra til å redusere ulikhet.

Table 6 viser at den økonomiske vekstvariabelen har en svak positiv sammenheng med ulikhet, som kan tyde på at økonomisk vekst på kort sikt ikke nødvendigvis er likt fordelt mellom regioner.

Variabelen andelen av befolkningen som er 65 år og eldre er ikke signifikant, men ifølge M. Dolls, K. Doorley, A. Paulus, H. Schneider, and E. Sommer [4] peker at population endringer, særlig aldring, kan bidra til økende ulikhet over tid.

P-verdien på 0.003 i Table 7 indikerer at modellen samlet sett er statistisk signifikant. Videre Table 7 viser R.squared 0.59 og justert R.squared 0.49. Der betyr R.squared at de fire varibalene, økonomisk utvikling, utdanning, transport og population forklarer om lag 59 % av ulikheten mellom regionene. Mens justert R.squared 0.49, som betyr at etter justering for antall variabler, modellen forklarer omtrent 49 % av variasjonen i Gini-koeffisienten. Modellen viser moderat

forklaringskraft. Dette indikerer at modellen fanger opp noen strukturelle påvirkningsfaktorer, men andre faktorer kan fortsatt bidra til regional ulikhet.

Modelltolkning

Forklar koeffisientene

Den multiple lineære regresjonsmodellen estimerer hvordan økonomisk utvikling, utdanning, transport og population påvirker regional ulikhet på NUTS2-nivå i 2017. I den multiple lineære regresjonsmodellen er utdanning den mest signifikante variabelen, veitetthet har en moderat signifikant effekt, økonomisk utvikling er svakt signifikant, mens population ikke er signifikant.

Koeffisienten for økonomisk utvikling (den prosentvise endringen av BNP per innbygger) er positiv og svakt signifikant (om lag p-verdien på 0.057). Dette innebærer at en økning på en prosent i BNP per innbygger i gjennomsnitt er assosiert med et svakt signifikant i Gini-koeffisienten, når de andre variablene holdes konstante. Dette stemmer med hypotesen i C. Lessmann and A. Seidel [1] om en N-formet sammenheng mellom utvikling og ulikhet, raskere økonomisk vekst kan bidra til økte regionale forskjeller.

Koeffisienten for utdanning (andelen av befolkningen med høyere utdanning) er positiv og statistisk signifikant på 1 %-nivået (om lag p-verdien på 0.01). Dette betyr at regioner med en høyere andel personer med høyere utdanning har lavere inntektsulikhet, alt annet likt. Dette stemmer overens med funnene fra D. Coady and A. Dizioli [2], som viser at høyere utdanningsnivå reduserer ulikhet.

Koeffisienten for transport (veitetthet) er negativ, med en p-verdi på om lag 0.039. En økning i veinettets tetthet er assosiert med lavere Gini-koeffisient, noe som indikerer at bedre transportinfrastruktur kan redusere ulikhet. Dette stemmer overens med funnene i C. Calderón and L. Servén [3], som viser at infrastruktur av høy kvalitet bidrar til å redusere inntektsforskjeller mellom regioner.

Koeffisienten for population (andelen av befolkningen som er 65 år og over) er negativ, men ikke signifikant (om lag p-verdi 0.516). Dette betyr at det ikke finnes tilstrekkelig statistisk grunnlag til å hevde at andelen eldre har en direkte effekt på ulikhet i denne modellen. M. Dolls, K. Doorley, A. Paulus, H. Schneider, and E. Sommer [4] påpeker imidlertid at demografiske endringer, særlig en økende andel eldre i befolkningen, over tid kan føre til høyere inntektsulikhet.

Samlet sett viser resultatene at utdanning har en sterk og signifikant positiv effekt på ulikhet, mens veitetthet har en moderat, men signifikant negativ effekt. Økonomisk vekst har en svakt signifikant positiv effekt på ulikhet, mens aldringseffekten er statistisk ubetydelig. Dette antyder at investeringer i utdanning og transportinfrastruktur kan være effektive virkemidler for å redusere regionale ulikheter.

Reflekter

Den multiple lineære regresjonsmodellen har en moderat forklaringskraft. Dette kan se fra justert R.squared på om lag 0.49. Det vil si at de fire variblene i modellen min forklarer rundt halvparten av variasjonen i regional ulikhet mellom NUTS2-regionene for de fire landene Belgia, Nederland, Bulgaria og Norge.

Utdanning på høyt nivå har størst betydning for å forklare ulikhet, og viser en sterk og signifikant positiv sammenheng. Dette betyr at regioner med høyere utdanningsnivå har større inntektsulikhet. Transport (veitetthet) bidrar også signifikant, men i motsatt retning, der høyere veinettstetthet reduserer ulikhet. Økonomisk utvikling har en svak positiv effekt og viser at rask vekst ikke nødvendigvis gir jevnere fordeling mellom regioner. Population (65 år og over) har en negativ koeffisient, men er statistisk ubetydelig i modellen min. Likevel kan dette ikke si at denne faktoren ikke har noen innvirkning på ulikhet, ettersom demografiske endringer kan ha effekter som ikke fanges opp av modellen.

Modellen har en tilfredsstillende, men moderat forklaringskraft og fanger opp sentrale faktorer som utdanning og transportinfrastruktur bak regional ulikhet. Selv om modellen forklarer en del av ulikheten, viser den også at forhold som ikke er inkludert i modellen, fortsatt kan ha viktige effekter. Derfor forklarer modellen ikke fullt ut variasjonen i ulikhet. For å oppnå en dypere forståelse av årsakene til ulikhet og øke modellens forklaringskraft, kreves det bredere og mer detaljerte analyser.

Del C: Dokumenter din bruk av AI

I denne A2-oppgaven brukte jeg OpenAI ChatGPT (GPT-5-modellen) som et støtteverktøy i arbeidsprosessen. Verktøyet ble brukt til å rette kodefeil, oversette språk, forstå fagbegreper og faglig kunnskap, finne og oppsummere litteratur, tolke datainnhold samt forbedre struktur og språklig klarhet i teksten, og øke kvaliteten på den skriftlige framstillingen.

ChatGPT hjalp meg med å hente, forstå og bearbeide ulike datasett fra Eurostat, spesielt i del B da jeg hentet nye datasett og skulle sammenstille og aggregere informasjon. Verktøyet utvidet perspektivet mitt på hvordan slike data kan struktureres og analyseres. Det hjalp meg også med å tolke hvordan variabler som veitetthet og andel personer med høyere utdanning påvirker regional ulikhet, samt med å forstå resultatene fra den multiple lineære regresjonsmodellen.

I tillegg bidro KI-verktøyet til å forbedre strukturen og den norske språkføringen i rapporten, slik at teksten ble mer presis og sammenhengende. Jeg brukte en trinnvis og målrettet spørrestrategi, der jeg kombinerte norsk og kinesisk for å få tydelige og presise uttrykk. Jeg kontrollerte og redigerte alltid innholdet for å forhindre overredigering, og for å sikre at alle analyser og konklusjoner bygget på min egen forståelse og mine data.

Bibliography

[1] C. Lessmann and A. Seidel, "Regional inequality, convergence, and its determinants – A view from outer space," *European Economic Review*, vol. 92, pp. 110–132, 2017, doi: https://doi.org/10.1016/j.euroecorev.2016.11.009.

[2] D. Coady and A. Dizioli, "Income Inequality and Education Revisited: Persistence, Endogeneity, and Heterogeneity," *IMF Working Papers*, vol. 17, no. 126, p. 1, 2017, doi: 10.5089/9781475595741.001.

[3] C. Calderón and L. Servén, "The Effects of Infrastructure Development on Growth and Income Distribution," Washington, D.C., 2004. [Online]. Available at: https://hdl.handle.net/10986/14136

[4] M. Dolls, K. Doorley, A. Paulus, H. Schneider, and E. Sommer, "Demographic change and the European income distribution," *The Journal of Economic Inequality*, vol. 17, no. 3, pp. 337–357, 2019.