

# Assignment2 msb104

Frida Alendal og Katinca Valvatne

## Table of contents

<b>1 Del A:</b>	<b>1</b>
1.1 Data og utvalg . . . . .	4
1.2 Modell . . . . .	5
1.3 Resultater og tolkning . . . . .	6
1.4 Slik leses figuren . . . . .	7
1.5 Hva det betyr i praksis . . . . .	8
<b>2 Del B: Andre drivkrefter bak regional ulikhet</b>	<b>8</b>
<b>3 Del C: Kildebruk</b>	<b>11</b>
3.1 Datakilder . . . . .	11
3.2 Bruk av KI-verktøy (I tråd med HVL sin bruk av KI) . . . . .	12

```
# Les alle .rds i data/ og gjør dem tilgjengelige som objekter
rds_files <- list.files("data", pattern = "\\rds$", full.names = TRUE)
stopifnot(length(rds_files) > 0) # vennlig feilmelding om mappa er tom

loaded <- lapply(rds_files, readRDS)
names(loaded) <- sub("\\rds$", "", basename(rds_files))
list2env(loaded, .GlobalEnv)
```

<environment: R\_GlobalEnv>

```
# Sjekk at nøkkelobjektene finnes
needed <- c("gini_nuts2_year", "dev_2000_2017")
missing <- setdiff(needed, ls())
if (length(missing)) stop("Mangler: ", paste(missing, collapse = ", "),
                        ". Lagre dem først i Ass1 (saveRDS).")
```

## 1 Del A:

```
# Pakker vi trenger -----
# dplyr/tidyr: datamanipulasjon, broom: pene modelltabeller, ggplot2: figur
library(dplyr)
library(tidyr)
library(broom)
library(ggplot2)
```

```
library(dplyr)
library(dineq) # gini.wtd

land <- c("ES","AT","CZ","SI","EE") # landene deres

# Sjekk at kolonnene finnes (gir klar feilmelding om noe mangler)
stopifnot(all(c("year","gdpc_eur","pop","NUTS2","country","geo") %in% names(gdpc_clean)))

# 1) Avgrens til riktige land og år (bruk YEAR, ikke time)
n3 <- gdpc_clean %>%
  filter(country %in% land, year >= 2000, year <= 2017) %>%
  select(country, NUTS2, geo, year, gdpc = gdpc_eur, pop)

# 2) Aggreger til NUTS-2 per år:
#   - mean_y: pop-vektet BNP p.c.
#   - gini_w: GINI innen NUTS-2 (beregnet fra NUTS-3)
gini_nuts2_year <- n3 %>%
  group_by(country, NUTS2, year) %>%
  summarise(
    n_nuts3 = dplyr::n_distinct(geo), # hvor mange NUTS-3 inngår
    mean_y = weighted.mean(gdpc, pop, na.rm = TRUE), # pop-vektet gj.snitt
    gini_w = dineq::gini.wtd(gdpc, weights = pop), # GINI av NUTS-3 verdier
    .groups = "drop"
  )
```

```
library(dplyr)
library(tidyr)

# (A) Avgrens til landene deres (hvis du vil)
land <- c("ES","AT","CZ","SI","EE")
g2 <- gini_nuts2_year %>% filter(country %in% land)

# (B) Finn "baseline"-år per NUTS-2: 2000 hvis finnes, ellers første år 2000
#   -> Sikrer at regioner som starter i f.eks. 2001 ikke droppes.
base_year <- g2 %>%
  group_by(NUTS2, country) %>%
  summarise(
    y0_year = ifelse(any(year == 2000),
                     2000,
                     min(year[year >= 2000], na.rm = TRUE)),
```

```

    .groups = "drop"
  )

# (C) Hent BNP p.c. i baseline-året (y0) og i 2017 (y17)
y0 <- g2 %>%
  inner_join(base_year, by = c("NUTS2", "country")) %>%
  filter(year == y0_year) %>%
  transmute(NUTS2, country, y0 = mean_y)

y17 <- g2 %>%
  filter(year == 2017) %>%
  transmute(NUTS2, country, y17 = mean_y)

# (D) Beregn utvikling som log-vekst: log(y17) - log(y0)
# -> Nå får du med alle regioner som har begge år (baseline og 2017).
dev_2000_2017 <- y0 %>%
  inner_join(y17, by = c("NUTS2", "country")) %>%
  mutate(dev_log = log(y17) - log(y0)) %>%
  select(country, NUTS2, dev_log)

# (E) Rask sanity check: hvor mange NUTS-2 fikk vi per land?
dev_2000_2017 %>%
  count(country, name = "n_nuts2")

```

```

# A tibble: 5 x 2
  country n_nuts2
  <chr>    <int>
1 AT         7
2 CZ         1
3 EE         1
4 ES        10
5 SI         2

```

```

# Lager kryss-snittet for 2017
# 1) GINI i 2017 (målt på NUTS-2). Vi trenger dette som Y-variabel.
gini_2017 <- gini_nuts2_year %>%
  dplyr::filter(year == 2017) %>%
  dplyr::select(country, NUTS2, gini_w, n_nuts3)

# 2) Slå sammen med utviklingsmålet (X-variabelen)
# -> cs_2017 blir datasettet vi estimerer på.
cs_2017 <- gini_2017 %>%
  dplyr::inner_join(dev_2000_2017, by = c("country", "NUTS2"))

# 3) Sjekk hvor mange regioner per land vi faktisk analyserer (din tabell)
coverage <- cs_2017 %>%

```

```
dplyr::count(country, name = "n_nuts2")
coverage
```

```
# A tibble: 5 x 2
  country n_nuts2
  <chr>    <int>
1 AT         7
2 CZ         1
3 EE         1
4 ES        10
5 SI         2
```

## 1.1 Data og utvalg

Vi analyserer NUTS-2-regioner i **AT, CZ, EE, ES og SI**. Avhengig variabel (**Y**) er **GINI i 2017** på NUTS-2, beregnet fra NUTS-3 med befolkningsvekter. Forklaringsvariabelen (**X**) er **regional utvikling** definert som log-vekst i BNP per innbygger fra første tilgjengelige år 2000 til 2017.

Vi inkluderer bare regioner som har både baseline og 2017. Dette gir totalt **21** NUTS-2 regioner. Dekning per land vises i Tabell ?@tbl-coverage.

```
# Regioner som mangler 2017 eller ikke har nok NUTS-3 til å beregne GINI
mangler_2017 <- gini_nuts2_year %>%
  dplyr::group_by(country, NUTS2) %>%
  dplyr::summarise(
    har_2017 = any(year == 2017),
    n_nuts3_2017 = dplyr::if_else(
      har_2017, max(n_nuts3[year == 2017], na.rm = TRUE), NA_real_,
      .groups = "drop"
    ) %>%
  dplyr::filter(!har_2017 | is.na(n_nuts3_2017) | n_nuts3_2017 < 2)

# Regioner som har 2017 men som mangler baseline-år for vekst
har_2017 <- gini_nuts2_year %>%
  dplyr::filter(year == 2017) %>%
  dplyr::select(country, NUTS2) %>%
  dplyr::distinct()

i_cs <- cs_2017 %>% dplyr::select(country, NUTS2) %>% dplyr::distinct()
mangler_i_cs <- dplyr::anti_join(har_2017, i_cs, by = c("country", "NUTS2"))

mangler_2017 # til appendiks/notat

# A tibble: 0 x 4
# i 4 variables: country <chr>, NUTS2 <chr>, har_2017 <lgl>, n_nuts3_2017 <dbl>
```

```
mangler_i_cs    # til appendiks/notat
```

```
# A tibble: 0 x 2
# i 2 variables: country <chr>, NUTS2 <chr>
```

## 1.2 Modell

Vi undersøker om regional økonomisk utvikling henger sammen med ulikhet mellom regioner i 2017. Ulikhet måles som Gini på NUTS-2-nivå, og utvikling måles som vekst i BNP per innbygger fram til 2017. Tanken er at regioner som har vokst mer, kan få større forskjeller internt, men dette trenger ikke være sterkt eller gjelde overalt.

Sterkere økonomisk opphenting kan øke intern spredning (positivt ), men sammenhengen kan være svak og påvirkes av regionale strukturer og sammensetning.

```
# Vis hva som faktisk er i kryss-snittet vårt -----
# (du bør allerede ha laget cs_2017 tidligere i dokumentet)
dplyr::glimpse(cs_2017)      # rask struktur
```

```
Rows: 21
Columns: 5
$ country <chr> "AT", "AT", "AT", "AT", "AT", "AT", "AT", "CZ", "EE", "ES", "E~
$ NUTS2    <chr> "AT11", "AT12", "AT21", "AT22", "AT31", "AT32", "AT33", "CZ05"~
$ gini_w   <dbl> 0.04327867, 0.13510424, 0.09620401, 0.12164526, 0.12201980, 0.~
$ n_nuts3  <int> 3, 7, 3, 6, 5, 3, 5, 3, 3, 4, 3, 3, 9, 5, 4, 3, 3, 8, 7, 8, 4
$ dev_log  <dbl> 0.4806890, 0.4609958, 0.4453069, 0.4778743, 0.4786247, 0.49265~
```

```
nrow(cs_2017)      # antall regioner
```

```
[1] 21
```

```
coverage <- cs_2017 %>%
  dplyr::count(country, name = "n_nuts2")
coverage      # dekning per land (skrives ut i pdf)
```

```
# A tibble: 5 x 2
  country n_nuts2
  <chr>    <int>
1 AT              7
2 CZ              1
3 EE              1
4 ES             10
5 SI              2
```

```
# OLS med robuste SE -----
m1 <- lm(gini_w ~ dev_log, data = cs_2017)

# Pene tabeller i dokumentet:
if (!requireNamespace("knitr", quietly = TRUE)) install.packages("knitr")
if (!requireNamespace("broom", quietly = TRUE)) install.packages("broom")
if (!requireNamespace("sandwich", quietly = TRUE)) install.packages("sandwich")
if (!requireNamespace("lmtest", quietly = TRUE)) install.packages("lmtest")

knitr::kable(broom::tidy(m1), caption = "OLS: GINI (2017) på utvikling (log-vekst)")
```

Table 1: OLS: GINI (2017) på utvikling (log-vekst)

term	estimate	std.error	statistic	p.value
(Intercept)	0.0208903	0.0206449	1.011887	0.3243057
dev_log	0.0860550	0.0370003	2.325793	0.0312527

```
knitr::kable(broom::glance(m1), caption = "Modellmål (R2, adj. R2, n)")
```

Table 2: Modellmål (R<sup>2</sup>, adj. R<sup>2</sup>, n)

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.2216086	0.1806406	0.0407445	4.409314	0.0312527	1	38.46254	-	-	0.0315414	19	21
						70.92509	67.79152				

```
# Robuste standardfeil (HC1)
lmtest::coeftest(m1, vcov = sandwich::vcovHC(m1, type = "HC1"))
```

t test of coefficients:

```

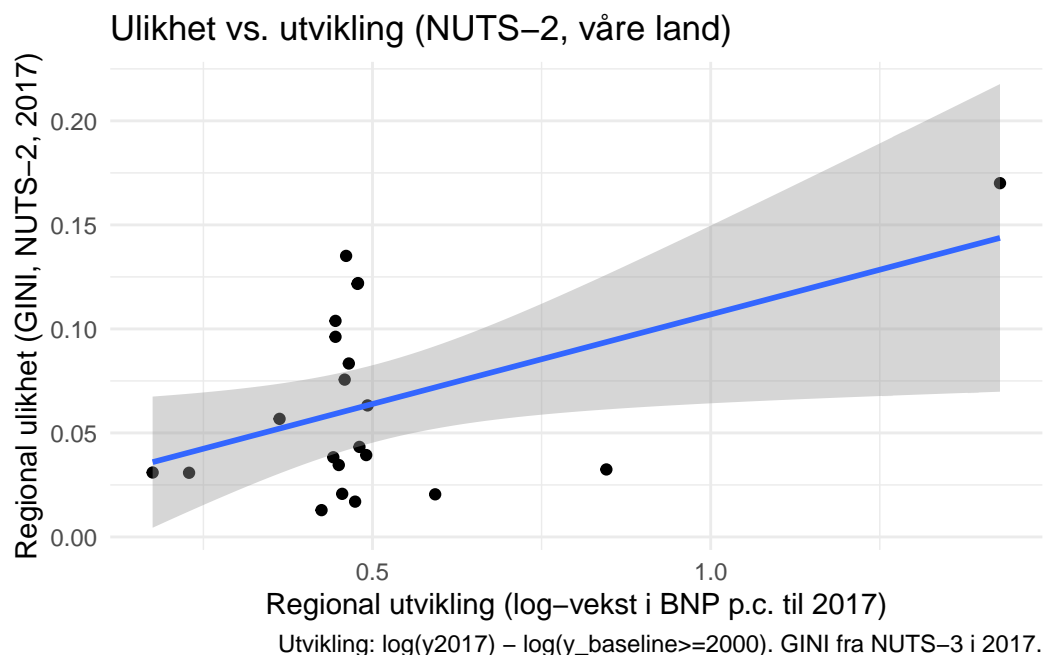
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.020890   0.015558  1.3428 0.195168
dev_log      0.086055   0.028888  2.9789 0.007713 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### 1.3 Resultater og tolkning

Modellen indikerer en positiv sammenheng mellom regional utvikling og ulikhet i 2017. Koeffisienten på utvikling er ... `t = r sprintf("%.2f", t_hc1)`, `p = r format.pval(p_hc1, digits = 3)` (`n = r nobs_m1`). Tolket i skala: rundt 10% høyere vekst ( $\Delta \log = 0.10$ ) er forbundet med `r sprintf("%.3f", 0.10*beta_hat)` høyere Gini i 2017. Modellen har moderat forklaringskraft ( $R^2 = r sprintf("%.2f",$

r2), justert  $R^2 = r$  `sprintf("%.2f", r2a))`, og spredningen i Figur ?@fig-ulikhet-utvikling viser at vekst alene ikke forklarer all variasjon mellom regioner.

```
# Figur -----
ggplot2::ggplot(cs_2017, ggplot2::aes(x = dev_log, y = gini_w, label = NUTS2)) +
  ggplot2::geom_point() +
  ggplot2::geom_smooth(method = "lm", se = TRUE) +
  ggplot2::labs(
    x = "Regional utvikling (log-vekst i BNP p.c. til 2017)",
    y = "Regional ulikhet (GINI, NUTS-2, 2017)",
    title = "Ulikhet vs. utvikling (NUTS-2, våre land)",
    caption = "Utvikling: log(y2017) - log(y_baseline 2000). GINI fra NUTS-3 i 2017."
  ) +
  ggplot2::theme_minimal()
```

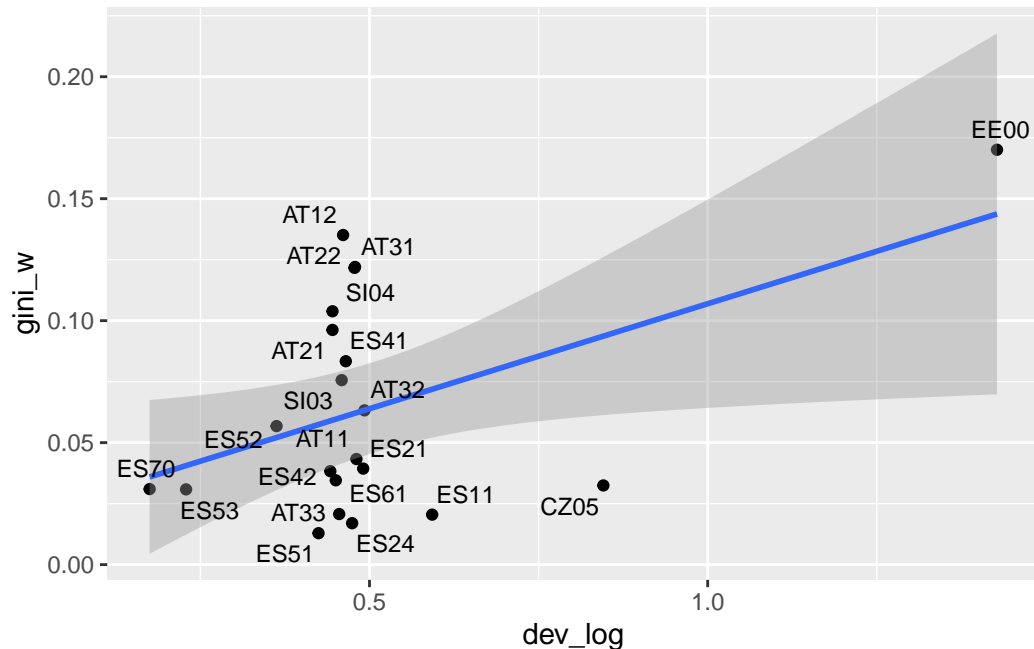


## 1.4 Slik leses figuren

Figuren viser sammenhengen mellom regional utvikling (x-akse: log-vekst i BNP p.c. fram til 2017) og ulikhet (y-akse: Gini i 2017 på NUTS-2). Hver prikk er én region, den blå linjen er OLS-estimatet, og det grå båndet er 95% konfidensintervall. Helningen er svak positiv, som betyr at regioner med høyere vekst tenderer til noe høyere intern ulikhet i 2017. Usikkerheten er relativt stor (spesielt ved høy vekst, der vi har færre observasjoner), så vekst alene forklarer ikke all variasjon mellom regionene.

```
# install.packages("ggrepel")
library(ggrepel)
ggplot(cs_2017, aes(dev_log, gini_w, label = NUTS2)) +
```

```
geom_point() +  
geom_smooth(method = "lm", se = TRUE) +  
ggrepel::geom_text_repel(size = 3)
```



Figuren over viser den samme sammenhengen mellom utvikling og ulikhet som i hovedfiguren, men her er hver prikk merket med NUTS-2-koden til regionen. Det gjør det lettere å se hvilke regioner som driver mønsteret i dataene. For eksempel ligger EE00 høyt både på vekst og ulikhet, mens flere spanske regioner (ES-kodene) har forholdsvis lav ulikhet gitt utviklingsnivået sitt. Poenget er ikke å tolke hver enkelt region i detalj, men å illustrere hvilke konkrete regioner som ligger nærmest og lengst fra regresjonslinjen.

## 1.5 Hva det betyr i praksis

Funnene peker på at sterk regional vekst kan følges av noe større forskjeller innad i regionen. For politikere og planleggere betyr det at tiltak som fremmer vekst kan med fordel kombineres med tiltak som sørger for bred deltakelse i verdiskapingen – for eksempel kompetanseheving, mobilitet i arbeidsmarkedet og inkluderende næringsutvikling.

## 2 Del B: Andre drivkrefter bak regional ulikhet

I del A så vi på sammenhengen mellom regionalt utviklingsnivå (log BNP per innbygger) og inntektsulikhet målt ved Gini for 21 NUTS2 regioner i Estland, Slovenia, Tsjekkia, Østerrike og Spania. Det ga et første-bilde av hvordan rikere regioner skiller seg fra fattigere regioner. I del B utvider vi denne modellen ved å legge til tre regionale kjennetegn i 2017: andel av befolkningen med høyere utdanning, arbeidsledighetsrate og befolkningstetthet. Variablene hentes fra Eurostat og kobles til



datasettet fra del A. Målet er å undersøke om disse faktorene kan bidra til å forklare hvorfor noen regioner har høyere inntektsulikhet enn andre, utover forskjeller i utviklingsnivå alene.

```
library(readxl)
library(dplyr)

# 1. Gini + evt. growth fra del A (bytt ut med deres objekt/fil)
gini_2017 <- gini_2017 # eller read_rds("gini_2017.rds") osv.

# 2. Utdanning
edu <- read_excel("edat_lfse_04_2017_tertiary_Y25_64.xlsx") |>
  rename(
    nuts2 = `GEO (Codes)`,
    edu_tertiary = `2017`
  ) |>
  filter(!is.na(edu_tertiary))

# 3. Arbeidsledighet
unemp <- read_excel("tgs00010_unemployment_2017.xlsx") |>
  rename(
    nuts2 = `GEO (Codes)`,
    unemp_rate = `2017`
  ) |>
  filter(!is.na(unemp_rate))

# 4. Befolkningstetthet
dens <- read_excel("tgs00024_popdensity_2017.xlsx") |>
  rename(
    nuts2 = `GEO (Codes)`,
    pop_density = `2017`
  ) |>
  filter(!is.na(pop_density))

# 5. slå sammen alt i ett datasett til del B
data_B <- gini_2017 |>
  left_join(edu, by = c("NUTS2" = "nuts2")) |>
  left_join(unemp, by = c("NUTS2" = "nuts2")) |>
  left_join(dens, by = c("NUTS2" = "nuts2"))
```

```
glimpse(data_B)
```

Rows: 21

Columns: 7

```
$ country    <chr> "AT", "AT", "AT", "AT", "AT", "AT", "AT", "CZ", "EE", "ES~
$ NUTS2      <chr> "AT11", "AT12", "AT21", "AT22", "AT31", "AT32", "AT33", "~
$ gini_w     <dbl> 0.04327867, 0.13510424, 0.09620401, 0.12164526, 0.1220198~
$ n_nuts3    <int> 3, 7, 3, 6, 5, 3, 5, 3, 3, 4, 3, 3, 9, 5, 4, 3, 3, 8, 7, ~
```

```
$ edu_tertiary <chr> "28", "32", "29.7", "27.8", "29.2", "31.2", "29", "20.3", ~
$ unemp_rate <chr> "5.2", "4.8", "4.8", "4.5", "4", "3.1", "3.3", "2.8", "5.~
$ pop_density <chr> "79.7", "88.2", "59.9", "76.2", "125.4", "78.099999999999~
```

```
nrow(data_B) #Bare for å sjekke
```

```
[1] 21
```

```
# Slå sammen utviklingsvariabelen inn i data_B
data_B2 <- data_B |>
  left_join(dev_2000_2017 |> select(NUTS2, dev_log),
            by = "NUTS2")

glimpse(data_B2)
```

```
Rows: 21
Columns: 8
$ country <chr> "AT", "AT", "AT", "AT", "AT", "AT", "AT", "CZ", "EE", "ES~
$ NUTS2 <chr> "AT11", "AT12", "AT21", "AT22", "AT31", "AT32", "AT33", "~
$ gini_w <dbl> 0.04327867, 0.13510424, 0.09620401, 0.12164526, 0.1220198~
$ n_nuts3 <int> 3, 7, 3, 6, 5, 3, 5, 3, 3, 4, 3, 3, 9, 5, 4, 3, 3, 8, 7, ~
$ edu_tertiary <chr> "28", "32", "29.7", "27.8", "29.2", "31.2", "29", "20.3",~
$ unemp_rate <chr> "5.2", "4.8", "4.8", "4.5", "4", "3.1", "3.3", "2.8", "5.~
$ pop_density <chr> "79.7", "88.2", "59.9", "76.2", "125.4", "78.099999999999~
$ dev_log <dbl> 0.4806890, 0.4609958, 0.4453069, 0.4778743, 0.4786247, 0.~
```

```
data_B3 <- data_B2 |>
  # Gjør utdanning, ledighet og tetthet om til tall
  dplyr::mutate(
    edu_tertiary = as.numeric(edu_tertiary),
    unemp_rate = as.numeric(unemp_rate),
    pop_density = as.numeric(pop_density)
  )
```

```
mod_B <- lm(
  # Linær regresjon for del B:
  # Gini forklart av utvikling + tre regionkjennetegn
  gini_w ~ dev_log + edu_tertiary + unemp_rate + pop_density,
  data = data_B3
)

summary(mod_B)
```

Call:

```
lm(formula = gini_w ~ dev_log + edu_tertiary + unemp_rate + pop_density,
    data = data_B3)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.058219	-0.030800	0.005233	0.026714	0.059263

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.800e-02	5.321e-02	0.902	0.380
dev_log	5.205e-02	4.403e-02	1.182	0.254
edu_tertiary	6.488e-04	1.687e-03	0.385	0.706
unemp_rate	-2.112e-03	1.424e-03	-1.484	0.157
pop_density	-7.684e-05	1.296e-04	-0.593	0.561

Residual standard error: 0.04087 on 16 degrees of freedom

Multiple R-squared: 0.3405, Adjusted R-squared: 0.1756

F-statistic: 2.065 on 4 and 16 DF, p-value: 0.1333

I del B utvider vi modellen fra del A ved å inkludere tre regionale kjennetegn på NUTS2-nivå i 2017: andel med høyere utdanning, arbeidsledighetsrate og befolkningstetthet, i tillegg regionalt utviklingsnivå (log BNP per innbygger). Den utvidede modellen forklarer om lag 34% av variasjonen i Gini-koeffisienten mellom regionene ( $R^2 = 0,34$ ), men den justerte  $R^2$ -verdien er bare 0,18. F-testene for hele modellen er ikke statistisk signifikant på 5-prosentnivå ( $p = 0,13$ ), og ingen av de enkelte forklaringsvariablene har signifikante koeffisienter. Koeffisientene peker likevel i noen retninger som er interessante. Høyer BNP per innbygger og større andel med høyere utdanning er svakt positivt assosiert med høyere ulikhet. Høyere arbeidsledighet og høyere befolkningstetthet er svakt negativt assosiert med ulikhet når vi holder de andre variablene faste. Siden vi bare har 21 regioner og stor usikkerhet rundt estimatene, bør disse sammenhengene tolkes som indikasjoner og ikke som sterke bevis for kausale effekter.

### 3 Del C: Kildebruk

I denne oppgaven bygger vi på hermoniserte, offisielle statistikkilder fra Eurostat og et Gini-datasett basert på EU-SILC levert som vi selv bearbeidet i Assignment 1. Nedenfor gir vi en kort oversikt over hvilke kilder som er brukt, hvordan de er koblet, og hvilke svakheter de kan ha.

#### 3.1 Datakilder

- Gini 2017 (EU-SILC, bearbeidet i Assignment 1) Avhengig variabel er Gini-koeffisienten for disponible husholdningsinntekter i 2017 for NUTS2-regioner. Datasettet er konstruert av oss i Assignment 1, på grunnlag av EU-SILC-data og veiledning i emnet. Mulige svakheter er utvalgsusikkerhet og at små regioner kan ha mer støy i estimatene.

- Regionalt BNP per innbygger (Eurostat, `nama_10r_3gdp`) Brukes til å beregne log BNP per innbygger i 2000 og 2017, samt vekst i utviklingsnivå. Dette er nasjonalt harmoniserte nasjonalregnskapsdata, men tallene kan være reviderte og sier lite om fordeling innad i regionene.
- Regional befolkning per 1. januar (Eurostat, `demo_r_pjanaggr3`) Ble brukt i Assingment 1 til å beregne befolkningsvekter og BNP per innbygger.
- Andel med høyere utdanning (Eurostat, arbeidskraftundersøkelser) Variabelen `edu_tertiary` måler prosentandel av befolkningen 25–64 år med utdanning på nivå ED5–8. Dette er survey-baserte estimater og kan være noe usikre for små regioner.
- Arbeidsledighetsrate (Eurostat, `tgs00010`) Måler registrert arbeidsledighet i prosent av arbeidsstyrken (15–74 år) på NUTS2-nivå. Tids- og definisjonsendringer kan påvirke sammenlignbarheten, men dataserien er harmonisert av Eurostat.
- Befolkningstetthet (Eurostat, `tgs00024`) Antall personer per km<sup>2</sup> i 2017. Denne brukes som et enkelt mål på hvor urbane regionene er. En svakhet er at tetthet ikke fanger opp alle sider ved urbanisering og infrastruktur.

### 3.2 Bruk av KI-verktøy (I tråd med HVL sin bruk av KI)

I oppgaven har vi brukt ChatGPT som støtteverktøy, i tråd med HVL sine retningslinjer for bruk av kunstig intelligens. KI-verktøyet ble brukt gjennom nettversjonen av ChatGPT (OpenAI, modell GPT-5.1 Thinking).

ChatGPT er brukt til å få forslag til og feilsøke R-kode. Alle valg av variabler, modellspesifikasjon, tolkning av resultater og konklusjoner er våre egne, basert på pensum og oppgaveteksten. Prompt-strategien har vært å stille konkrete, avgrensede spørsmål, for eksempel om en bestemt feilmelding i R. Svarene er blitt vurdert kritisk og testet mot egne resultater før de eventuelt er tatt i bruk.