STK-IN1050 - Statistikk for informatikere

18. Lineær regresjon, del II

Matematisk institutt

2026-04-22

Multippel lineær regresjon

En lineær regresjonsmodell er gitt ved

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{ip} + \varepsilon_i \\ &= \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \qquad\qquad i = 1, \ldots, n, \end{aligned}$$

for $\varepsilon_i \sim N(0, \sigma^2)$ hvor $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ og $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$.

Vi gjør fire antakelser:

1. **Linearitet**: sammenhengen mellom responsen or forklaringsvariablene er (tilnærmet) lineær
2. **Uavhengige feil**: Feilleddene er uavhengig av hverandre
3. **Konstant varians**: Feilleddene har samme varians for alle verdier av $\boldsymbol{x}$
4. **Normalfordeling**: Feilleddene er (tilnærmet) normalfordelte

Minste kvadraters kriterium

For å konstruere en estimator for $\boldsymbol{\beta}$ bruker vi minste kvadraters kriteriet

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{b}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

hvor $\hat{y}_i = \boldsymbol{x}_i^T \boldsymbol{b} = b_0 + \sum_{j=1}^{p} b_j x_{ij}$.

Når vi har estimert regresjonskoeffisientene kan vi beregne **residulaene**

$$e_i = y_i - \hat{y}_i, \qquad i = 1, 2, \ldots, n.$$

Åpne spørsmål

- Hva er egenskapene til $\hat{\boldsymbol{\beta}}$?
- Hvordan kan vi estimere variansen σ^2 til feilleddene ε_i?
- Hvor stor er usikkerheten knyttet til estimatorene?
- Er det avhengighet mellom estimatorene?
- Hvordan kan vi lage konfidensintervall og gjøre hypotesetester for $\boldsymbol{\beta}$?
- **Hvordan kan vi bruke ferdig implementerte funksjoner til slik analyse?**

Vi bruker:

- `pandas` til å lese og håndtere data
- `statsmodels` til å estimere regresjonsmodellen
- `matplotlib` til visualisering

Hva kan forklare forskjeller i forventet levetid?

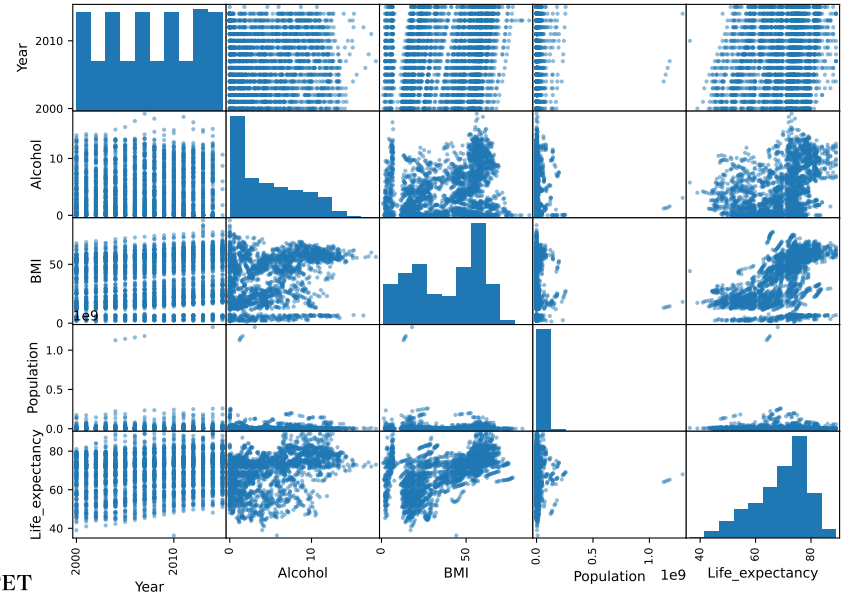
Data: forventet levetid (`Life_expectancy`) i ulike land, med forklaringsvariabler:

- `Year`: observasjonsår
- `Alcohol`: forbruk per innbygger (15+)
- `BMI` = body mass index
- `Population`: antall innbyggere

Vi antar en lineær sammenheng (multippel lineær regresjon):

$$y_i = \beta_0 + \beta_1 \cdot \text{Year}_i + \beta_2 \cdot \text{Alcohol}_i + \beta_3 \cdot \text{BMI}_i + \beta_4 \cdot \text{Population}_i + \varepsilon_i$$

hvor y_i er forventet levetid for observasjon i.


Year
Alcohol
BMI
Population
1e9
Life_expectancy
2000
2010
0
10
50
0.0
0.5
1.0
40
60
80

```
import pandas as pd
import numpy as np
import statsmodels.api as sm

# Read in data
LFdata = pd.read_csv("../data/Life_Expectancy_Data.csv", sep=",")
# Make names somewhat simpler
LFdata.columns = LFdata.columns.str.strip().str.replace(" ", "_")

# Extract y
LF_y = LFdata['Life_expectancy']
# Build X
colsX = ['Year','Alcohol','BMI','Population']
LF_X = sm.add_constant(LFdata[colsX])

# Fit model
LF_model = sm.OLS(LF_y, LF_X, missing="drop").fit()

# Print results
smry = LF_model.summary()
smry.extra_txt = ""   ## Remove footnotes
print(smry)
```

OLS Regression Results

Dep. Variable:	Life_expectancy	R-squared:	0.415
Model:	OLS	Adj. R-squared:	0.414
Method:	Least Squares	F-statistic:	374.5
Date:	Wed, 22 Apr 2026	Prob (F-statistic):	7.12e-244
Time:	09:18:45	Log-Likelihood:	-7271.5
No. Observations:	2114	AIC:	1.455e+04
Df Residuals:	2109	BIC:	1.458e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-559.2346	76.906	-7.272	0.000	-710.055	-408.414
Year	0.3069	0.038	8.008	0.000	0.232	0.382
Alcohol	0.5898	0.044	13.304	0.000	0.503	0.677
BMI	0.2434	0.009	26.625	0.000	0.226	0.261
Population	3.145e-09	2.62e-09	1.202	0.230	-1.99e-09	8.28e-09

Omnibus:	42.270	Durbin-Watson:	0.391
Prob(Omnibus):	0.000	Jarque-Bera (JB):	47.025
Skew:	-0.306	Prob(JB):	6.15e-11
Kurtosis:	3.399	Cond. No.	3.01e+10

Dagens mål

- Kunne lese og forstå output fra `statsmodels.OLS(...).fit().summary()`
- Kunne koble informasjon i output-tabellen til den underliggende teorien
- Kunne tolke en enkel multippel regresjonsmodell

Dagens mål

Output-tabellen har tre deler:

1. **Modellkvalitet** (R^2, justert R^2, F-test, AIC/BIC)
2. **Koeffisient-tabellen** (estimat, standardfeil, *t*, *p*, konfidensintervall)
3. **Diagnose-statistikk** (residualer, normalitet, Durbin–Watson, osv.)

Vi går igjennom disse delene en ad gangen og diskuterer et par ting ikke inkludert i tabellen.

Modellkvalitet

Toppen av tabellen ser typisk slik ut:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life_expectancy   R-squared:                       0.415
Model:                            OLS   Adj. R-squared:                  0.414
Method:                 Least Squares   F-statistic:                     374.5
Date:                             ...   Prob (F-statistic):          7.12e-244
Time:                             ...   Log-Likelihood:                -7271.5
No. Observations:                   n   AIC:                         1.455e+04
Df Residuals:                 n-(p+1)   BIC:                         1.458e+04
Df Model:                           p
Covariance Type:            nonrobust
```

Responsvariabel og modell

- **Dep. Variable:** `Life_expectancy` $\Rightarrow$ responsvariablen
- **Model: OLS** $\Rightarrow$ «Ordinary Least Squares» = minste kvadraters metode
  - Vi velger koeffisienter $\beta_0, \ldots, \beta_p$ som minimerer

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})\right)^2$$

- **Method: Least Squares** $\Rightarrow$ modellen er estimert med minste kvadraters metode

Antall observasjoner og frihetsgrader

- **No. Observations:** n
  - Hvor mange rader (observasjoner) som faktisk brukes (etter `missing="drop"`)
  - I vårt tilfelle: 2114
- **Df Model:** p
  - Antall forklaringsvariabler (uten konstant-leddet)
  - I vårt tilfelle: 4 (`Year`, `Alcohol`, `BMI`, `Population`)
- **Df Residuals:** $n - (p + 1)$
  - Antall frihetsgrader i residualene
  - $p + 1$ fordi vi også estimerer konstantleddet β_0
  - I vårt tilfelle: $n = 2114$ og $p = 4 \Rightarrow$ Df Residuals $= 2114 - (4 + 1) = 2109$

R-kvadrat (determinasjonskoeffisienten)

R-kvadrat (R^2) måler hvor stor andel av variasjonen i y som modellen forklarer. Først definerer vi noen størrelser:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ Variabiliteten i y_i-ene

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$ Variabiliteten i $\hat{y}_i$-ene

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ Variabiliteten i residualene

Det kan vises at $SST = SSR + SSE$.

R-kvadrat (determinasjonskoeffisienten)

Vi kan nå definere R^2:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

dvs. forholdet mellom forklart varians (variabiliteten i $\hat{y}_i$-ene) og total varians (variabiliteten i y_i-ene).

Vi har at R^2 ligger mellom 0 og 1:

- Nær 1: modellen forklarer mye
- Nær 0: modellen forklarer lite

Justert R-squared

- **Adj. R-squared** korrigerer R^2 for antall forklaringsvariabler
- Idé: R^2 øker (eller blir uendret) når vi legger til flere variabler, selv om de ikke er nyttige
- Justert R^2 straffer «unødvendige» variabler og kan gå ned når vi legger til en dårlig forklaringsvariabel

$$R^2_{\text{adj}} = 1 - (1 - R^2) \cdot \frac{n - 1}{n - p - 1}$$

Her:

- n: antall observasjoner
- p: antall forklaringsvariabler (utenom konstant-leddet)

Log-Likelihood, AIC og BIC

- **Log-Likelihood**: logaritmen til sannsynligheten for å observere dataene gitt modellen: $\log(L)$
  - Basert på antakelse om normalfordelte feil: $y_i \sim N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_n x_{ip}, \sigma^2)$
- **AIC (Akaike Information Criterion)**: $\text{AIC} = 2k - 2\log(L)$
- **BIC (Bayesian Information Criterion)**: $\text{BIC} = k\log(n) - 2\log(L)$

Her:

- k: antall estimerte parametere ($k = p + 1$)
- n: antall observasjoner

Lavere AIC/BIC er «bedre» (balanse mellom god passform og enkel modell)

F-test

- En F-test tester «hele modellen» samlet:
  - H_0: Alle forklaringsvariabler har koeffisient 0 ($\beta_1 = \beta_2 = \cdots = \beta_p = 0$)
  - H_a: Minst én $\beta_j \neq 0$
- **F-statistic**: Testobservatoren for F-testen:

$$F = \frac{\text{SSR}/p}{\text{SSE}/(n-p-1)}$$

- **Prob(F-statistic)**: p-verdi for F-testen
  - Liten p-verdi: modellen som helhet har forklaringskraft
  - Stor p-verdi: ingen bevis for at modellen er bedre enn å bare bruke gjennomsnittet

Modellen i vektorform

Det er nyttig å skrive modellen kompakt:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\mathbf{y}$: $n \times 1$ vektor (observasjoner av levetid)
- $\mathbf{X}$: $n \times (p + 1)$ matrise
  - Første kolonne: 1 (konstant)
  - Resten: Year, Alcohol, BMI, Population
- $\boldsymbol{\beta}$: $(p + 1) \times 1$ vektor med koeffisienter
- $\boldsymbol{\epsilon}$: feilledd, $n \times 1$ vektor

Estimator for regresjonskoeffisiente

OLS-løsningsformelen:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Vi setter inn $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}.$$

Det følger at estimatoren $\hat{\boldsymbol{\beta}}$ er normalfordelt og estimeringsfeilen er gitt ved:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}.$$

Estimator for regresjonskoeffisiente

Variansen til estimatoren blir:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}\big((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\big) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{Var}(\boldsymbol{\epsilon})\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

Vi har at $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Dvs., hvis alle modellantakelsene stemmer:

$$\hat{\boldsymbol{\beta}} \sim N\big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\big).$$

Estimering av σ^2

Vi trenger også å estimere variansen til feilledene $\sigma^2 = \text{Var}[\varepsilon_i]$.

Idé: bruk residualer som estimat for feilene
$\varepsilon_i = y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j}$

Residualer:

$$\hat{\varepsilon}_i = e_i = y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{i,j}$$

Naivt variansestimat:

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \right)^2$$

Estimering av σ^2

Den naive estimatoren er forventingsskjev:

$$\mathbb{E}[\tilde{\sigma}^2] = \frac{n-(p+1)}{n}\sigma^2.$$

Dvs. $\tilde{\sigma}^2$ undervurderer σ^2 (særlig for små n).

En forventningsrett estimator er:

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{i,j} \right)^2.$$

Estimering av σ^2

`summary()` skriver ikke ut estimatet til σ^2, men det er inkludert i `LF_model` objektet vårt:

```
print(f"Sigma^2-hat: {LF_model.mse_resid:.3f}")
print(f"Sigma-hat: {np.sqrt(LF_model.mse_resid):.3f}")
```

```
Sigma^2-hat: 57.044
Sigma-hat: 7.553
```

Koeffisient-tabellen

	coef	std err	t	P>|t|	[0.025	0.975]
const	-559.2346	76.906	-7.272	0.000	-710.055	-408.414
Year	0.4069	0.038	8.008	0.000	0.232	0.382
Alcohol	0.5898	0.044	13.304	0.000	0.503	0.677
BMI	0.2434	0.009	26.625	0.000	0.226	0.261
Population	3.145e-09	2.62e-09	1.202	0.230	-1.99e-09	8.28e-09

Vi skal forklare:

- `coef`
- `std err`
- `t`
- `P>|t|`
- Konfidensintervallene `[0.025, 0.975]`

Koeffisienter (`coef`)

- `coef`-kolonnen viser $\hat{\beta}_j$ for hver forklaringsvariabel
- Tolkning:
  - **Year**: hvor mye forventet levealder endrer seg når år øker med 1, mens Alcohol, BMI og Population holdes konstant
  - **Alcohol**: endring i levealder for 1 enhet økning i alkoholkonsum
  - osv.
- Eksempel:
  - `Alcohol`-koeffisient 0.6 betyr: «Hvis alkoholkonsum øker med 1 enhet, forventes levealderen å øke med ca 0.60 år, alt annet likt.»

Konstantleddet (`const`)

- `const`: $\hat{\beta}_0$ i modellen

  $\hat{\beta}_0$ = forventet levealder når $\text{Year} = 0, \text{Alcohol} = 0, \text{BMI} = 0, \text{Population} = 0$

- Ofte er dette punktet utenfor det realistiske området: Derfor er konstantleddet ofte mindre interessant å tolke direkte
- Likevel viktig for prediksjoner (bestemmer nivået på regresjonslinjen)

Standardfeil (`std err`)

- Standardfeilen til $\hat{\beta}_j$ gir et mål på usikkerheten i estimatet
- Under antakelsene får vi:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

- Standardfeilen til $\hat{\beta}_j$ er kvadratroten av det j-te diagonalelementet
- I praksis estimerer vi σ^2 med residualene:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

t-statistikk (t) for hver koeffisient

- Vi ønsker å teste om hver variabel har effekt gitt at alle de andre variablene er inkludert i modellen:
  - Nullhypotese: $H_0 : \beta_j = 0$
  - Alternativ hypotese: $H_a : \beta_j \neq 0$
- t-statistikk:

$$t_j = \frac{\hat{\beta}_j - 0}{\text{SE}(\hat{\beta}_j)}$$

- Under H_0 og modell-antakelsene: t_j følger tilnærmet en t-fordeling med $n - p - 1$ frihetsgrader
- Stor absoluttverdi på $t_j \Rightarrow$ mer evidens mot H_0

P-verdi (`P>|t|`)

- P-verdi: sannsynlighet for å observere en t-statistikk minst like ekstrem som den vi fikk, hvis H_0 er sann
- For tosidig test:

$$p\text{-verdi} = 2 \cdot P\left(T_{n-p-1} \geq |t_j|\right)$$

- Tolkning:
  - Liten p-verdi (f.eks. < 0.05): variabelen har signifikant effekt
  - Stor p-verdi: vi har ikke bevis for at koeffisienten er forskjellig fra 0

Konfidensintervallene `[0.025  0.975]`

- 95% konfidensintervall for β_j:

$$\hat{\beta}_j \pm t_{0.975,\, n-p-1} \cdot \mathrm{SE}(\hat{\beta}_j)$$

- `statsmodels` rapporterer disse som kolonnene `[0.025, 0.975]`
- Tolkning:
  - Hvis vi gjentar eksperimentet mange ganger, da vil ca. 95% av slike intervaller inneholde den sanne β_j
  - Hvis intervallet ikke inneholder 0 $\Longrightarrow$ koeffisienten er signifikant på 5%-nivå

Skalering og tolkning

- Variabler på veldig ulike skalaer (f.eks. Population vs BMI) gir:
  - veldig små koeffisienter (som `-1.2e-08`)
  - vanskeligere direkte tolkning

- Mulige grep:
  - dele Population på 10^6 (millioner) eller 10^7
  - standardisere variabler

- Men: `summary()` viser alltid resultatene på den skalaen dataene har

Modelldiagnostikk

Omnibus:	42.270	Durbin-Watson:	0.391
Prob(Omnibus):	0.000	Jarque-Bera (JB):	47.025
Skew:	-0.306	Prob(JB):	6.15e-11
Kurtosis:	3.399	Cond. No.	3.01e+10

Dette er forskjellige diagnostikkmål:

- Omnibus, JB, Skew, Kurtosis: vurdering av normalitet
- Durbin–Watson: vurdering av korrelasjon i residualene
- Condition Number: handler om multikollinearitet (korrelasjon mellom forklaringsvariabler)

Normalitetstester: Omnibus, JB, Skew, Kurtosis

- **Skew** (skjevhet): Måler asymmetri i residualenes fordeling (0 = symmetrisk)

$$\text{Skew} = \frac{\frac{1}{n}\sum_{i=1}^{n}(e_i - \overline{e})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(e_i - \overline{e})^2\right)^{3/2}}$$

- **Kurtosis** (spisshet): Normalfordeling har kurtosis ca. 3

$$\text{Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(e_i - \overline{e})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(e_i - \overline{e})^2\right)^{2}}$$

e_i: residual for observasjon i, dvs. $e_i = y_i - \hat{y}_i$.

Normalitetstester: Omnibus, JB, Skew, Kurtosis

- **Omnibus** og **Jarque–Bera (JB)**-tester:
  - Nullhypotese: residualene er normalfordelte
  - Tester: kombinerer informasjon om skjevhet og kurtosis
  - `Prob(Omnibus)` og `Prob(JB)`: *p*-verdier
    - Liten *p*-verdi: residualene avviker fra normalfordeling
- Normalitetsantakelsen er viktig for *t*-tester og *F*-tester, spesielt ved små *n*

Durbin–Watson (DW)

- **DW-statistikk** sjekker for autokorrelasjon mellom residualer:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Autokorrelasjon: korrelasjon mellom e_i og $e_{i-1}, e_{i-2}, \ldots$

- Tolkning:
  - Verdi nær 2: ingen autokorrelasjon
  - Verdi < 2: positiv autokorrelasjon
  - Verdi > 2: negativ autokorrelasjon

Viktig for tidsserier (her: `Year` kan gi tidsstruktur)

Condition Number (Cond. No.)

- **Condition Number** måler hvor nær multikollinearitet vi er, dvs. hvor sterkt korrelerte forklaringsvariablene er
- Høye verdier (f.eks. > 30, > 100, > 1000) kan indikere problemer:
  - ustabile estimater
  - store standardfeil
- Årsak: Noen kombinasjoner av variabler gir nesten samme informasjon
- Typiske tiltak:
  - fjerne sterkt korrelerte variabler
  - kombinere variabler
  - bruke regularisering (ikke pensum)

Grafisk diagnostikk

I tillegg til det som skrives ut i `summary()`, kan det være nyttig å se på grafiske diagnostiker:

- Histogrammer og QQ-plots av residular $\Longrightarrow$ er residualene normalfordelte?
- Residular mot estimerte verdier $\Longrightarrow$ er det noen struktur her?
- Residular mot forklaringsvariabler $\Longrightarrow$ er det noen struktur her?

```
import numpy as np
import statsmodels.api as sm

np.random.seed(60)
N = 50
sig = 0.05
xsim = np.linspace(0, 10, N)
Xsim = sm.add_constant(xsim)

ysim1 = 1 + 0.5*xsim + np.random.normal(0, 0.5, N)
mod1 = sm.OLS(ysim1,Xsim).fit()
fit1 = mod1.fittedvalues
res1 = mod1.resid

ysim2 = 1 + 0.5*(xsim - 5)**2 + np.random.normal(0, 2, N)
mod2 = sm.OLS(ysim2,Xsim).fit()
fit2 = mod2.fittedvalues
res2 = mod2.resid

ysim3 = 1 + 0.5*xsim + (1 + xsim)*np.random.normal(0, 1, N)
mod3 = sm.OLS(ysim3,Xsim).fit()
fit3 = mod3.fittedvalues
res3 = mod3.resid
```

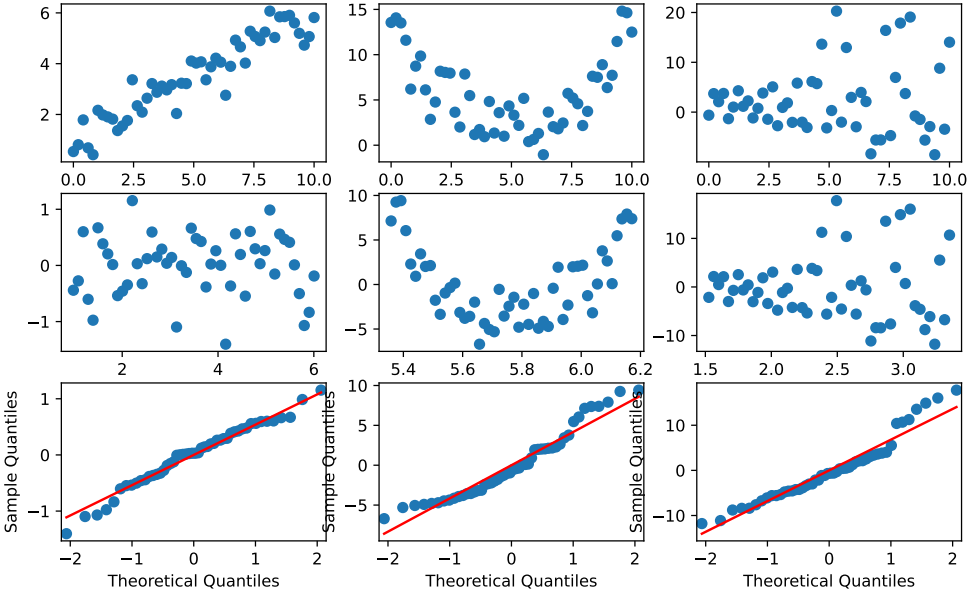
Vi plotter dataene, estimater mot residuler og QQ-plots av residualer:

```
axes[0,0].scatter(xsim, ysim1)
axes[0,1].scatter(xsim, ysim2)
axes[0,2].scatter(xsim, ysim3)

axes[1,0].scatter(fit1, res1)
axes[1,1].scatter(fit2, res2)
axes[1,2].scatter(fit3, res3)

sm.qqplot(res1, line='s', ax=axes[2,0])
sm.qqplot(res2, line='s', ax=axes[2,1])
sm.qqplot(res3, line='s', ax=axes[2,2])

plt.show()
```


Sample Quantiles
Theoretical Quantiles
Sample Quantiles
Theoretical Quantiles
Sample Quantiles
Theoretical Quantiles

Grafisk diagnostikk for levetidsdata

```
resid = LF_model.resid
fitted = LF_model.fittedvalues

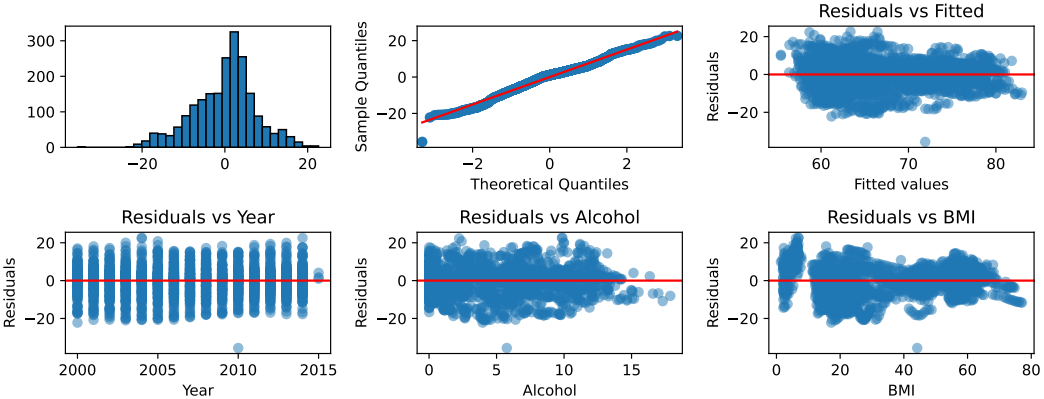
# Histogramm av residualene
axes[0].hist(resid,bins=30,edgecolor="black")

# QQ-plot av residualene mot normalfordeling
sm.qqplot(resid,line='s',ax=axes[1])

# Observasjoner mot estimater
axes[2].scatter(fitted, resid, alpha=0.5)

# Residualer mot enkelte forklaringsvariabler
axes[3].scatter(LFdata.loc[LF_model.resid.index,"Year"], resid, alpha=0.5)
axes[4].scatter(LFdata.loc[LF_model.resid.index,"Alcohol"], resid, alpha=0.5)
axes[5].scatter(LFdata.loc[LF_model.resid.index,"BMI"], resid, alpha=0.5)
```

Grafisk diagnostikk for levetidsdata



Oppsummering – tolkning av `summary()`

Når du ser på `LF_model.summary()`, tenk:

1. **Hva er modellen?**
   - Hvilke variabler er med, hvor mange observasjoner?
2. **Hvor god er passformen?**
   - R^2, justert R^2, F-statistic, Prob(F)
3. **Hvilke variabler er viktige?**
   - `coef`, `std err`, `t`, `P>|t|`, konfidensintervall
4. **Er antakelsene rimelige?**
   - Normalitet (Omnibus, JB, Q-Q-plot)
   - Uavhengighet (Durbin–Watson, kontekst)
   - Konstant varians (residual-plott)
   - Multikollinearitet (Cond. No., korrelasjonsmatrise)

De spørsmålene vi hadde i starten

- Hva er egenskapene til $\hat{\boldsymbol{\beta}}$? ✓
- Hvordan kan vi estimere variansen σ^2 til feilleddene ε_i? ✓
- Hvor stor er usikkerheten knyttet til estimatorene? ✓
- Er det avhengighet mellom estimatorene? ✓
- Hvordan kan vi lage konfidensintervall og gjøre hypotesetester for $\boldsymbol{\beta}$? ✓
- Hvordan kan vi bruke ferdig implementerte funksjoner til slik analyse? ✓