STK-IN1050 - Statistikk for informatikere

19. Lineær regresjon, del III

Matematisk institutt

2026-04-27

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 og 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**

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

Hva kan forklare forskjeller i forventet levetid?

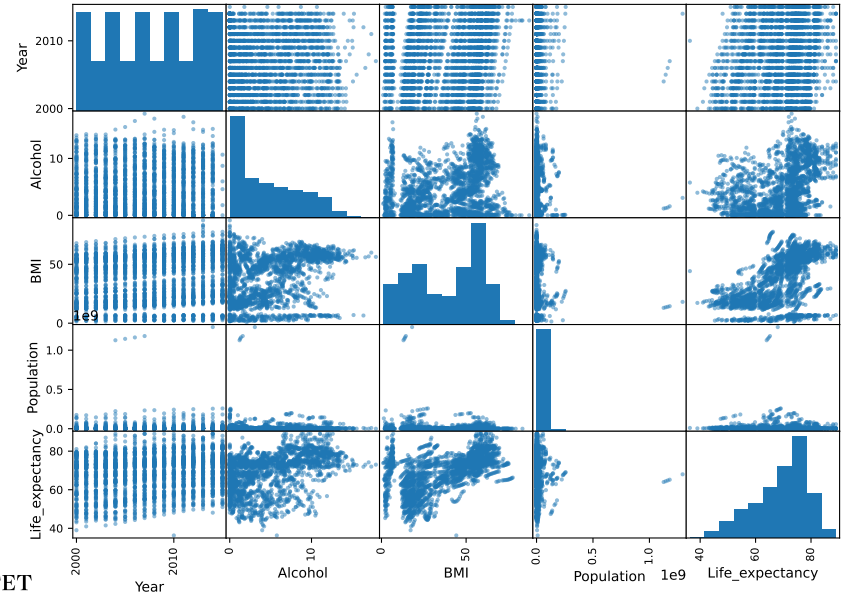
Data: forventet levetid (`Life_expectancy`) i ulike land, med forklaringsvariabler:

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

```
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:	Sun, 26 Apr 2026	Prob (F-statistic):	7.12e-244
Time:	19:11:17	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

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
```

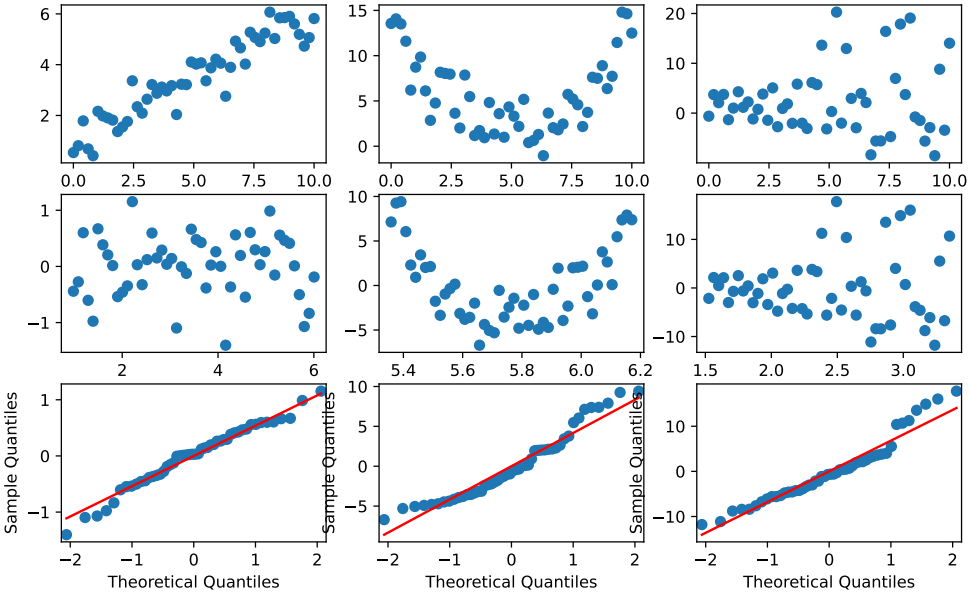
Vi plotter dataene, estimater mot residuler og QQ-plots av residualer:

```
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

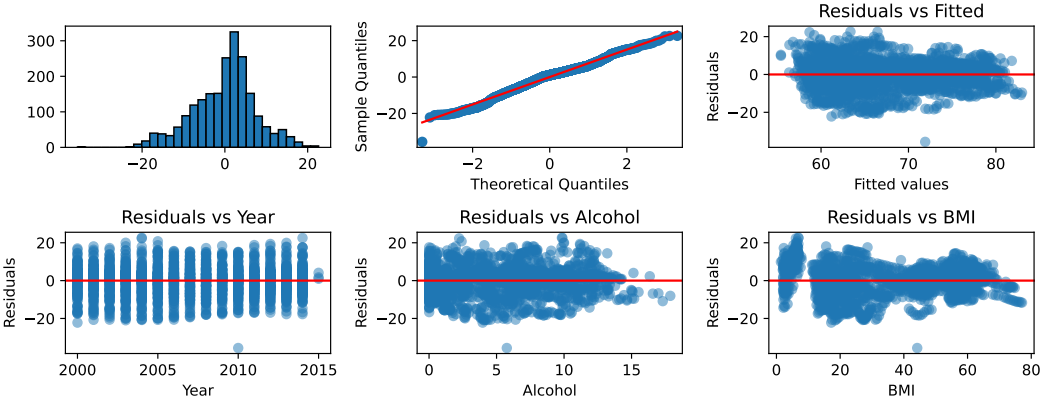
# Histogramm av residualene
axes[0].hist(resid,bins=30,edgecolor="black")

# 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), AIC, BIC
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)

Spørsmålene fra siste forelesning

- 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? ✓

To typer usikkerhet i lineær regresjon

Vi har til nå fokusert på estimering av sammenhengen mellom responsvariabelen y og forklaringsvariablene $x_1, \ldots, x_p$.

Vi skal nå vurdere to typer usikkerhet:

- **Estimeringsusikkerhet**: hvor sikre er vi på den sanne regresjonslinjen (det sanne forventningsnivået)?
- **Prediksjonsusikkerhet**: hvor sikre er vi på en prediksjon $\hat{y}$ av en ny observasjon?

Vi lager:

- **Konfidensintervall** for regresjonslinjen
- **Prediksjonsintervall** for en ny observasjon

Modellen og to typer intervaller

Vi bruker en standard lineær regresjonsmodell:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

For en gitt vektor $\mathbf{x}^*$ (konkrete verdier for Year, Alcohol, BMI, Population) har vi:

- **Forventet verdi**:
$$\mu(\mathbf{x}^*) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}^*) = (\mathbf{x}^*)^\top \boldsymbol{\beta}$$
- **Ny observasjon**:
$$\hat{y} = \mu(\mathbf{x}^*) + \varepsilon$$

Vi lager

- konfidensintervall for $\mu(\mathbf{x}^*)$
- prediksjonsintervall for $\hat{y}$

Hvorfor to intervaller?

Estimeringsusikkerhet:

- Vi bruker data til å estimere $\hat{\boldsymbol{\beta}}$.
- Dermed er også $\hat{\mu}(\mathbf{x}^*) = (\mathbf{x}^*)^\top \hat{\boldsymbol{\beta}}$ usikker.
- Konfidensintervallet for $\mu(\mathbf{x}^*)$ handler om å vurdere hvor den sanne regresjonslinjen ligger.

Prediksjonsusikkerhet:

- En ny observasjon $\hat{y}$ har både
  - usikkerhet i $\hat{\mu}(\mathbf{x}^*)$ (estimeringsusikkerhet), og
  - variasjon rundt regresjonslinjen (et tilfeldig avvik ε)
- Prediksjonsintervallet blir derfor **bredere** enn konfidensintervallet.

Eksempel: Forventet levetid

Vi velger et eksempelpunkt gitt ved gjennomsnittsverdier av alle forklaringsvariablene:

```
x_mean = LF_X.mean()
x_mean
```

```
const          1.000000e+00
Year           2.007519e+03
Alcohol        4.602861e+00
BMI            3.832125e+01
Population     1.275338e+07
dtype: float64
```

Dette gir en «typisk» kombinasjon av Year, Alcohol, BMI, Population.

Eksempel: Forventet levetid

Statsmodels har en to-delt metode: `get_prediction` + `summary_frame`.

```
# Create a data frame with the x-values we want to predict for
new_X = x_mean.to_frame().T  # row of mean values

pred = LF_model.get_prediction(new_X)
pred_summary = pred.summary_frame(alpha=0.05).round(3)
pred_summary
```

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	69.045	0.166	68.72	69.37	54.23	83.86

Eksempel: Forventet levetid

A $(1-\alpha)100\%$ **konfidensintervall** for $\mu(\mathbf{x}^*) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}^*)$ er gitt ved

$$\hat{\mu}(\mathbf{x}^*) \ \pm \ t_{n-p,\,\alpha/2} \times SE(\hat{\mu}(\mathbf{x}^*)) = 69.045 \ \pm \ 1.961 \times 0.166.$$

A $(1-\alpha)100\%$ **prediksjonsintervall** for $\hat{y}$ er gitt ved

$$\hat{\mu}(\mathbf{x}^*) \ \pm \ t_{n-p,\,\alpha/2} \cdot \sqrt{\hat{\sigma}^2 + SE(\hat{\mu}(\mathbf{x}^*))^2} = 69.045 \ \pm \ 1.961\sqrt{57.044 + 0.166^2}.$$

```
from scipy import stats
sigma2_hat = LF_model.mse_resid
t_crit = stats.t.ppf(1 - 0.05/2, df=LF_model.df_resid)
print(f"sigma2_hat = {sigma2_hat:.3f}, t_crit = {t_crit:.3f}")
```

```
sigma2_hat = 57.044, t_crit = 1.961
```

Visualisering – konfidens vs prediksjon

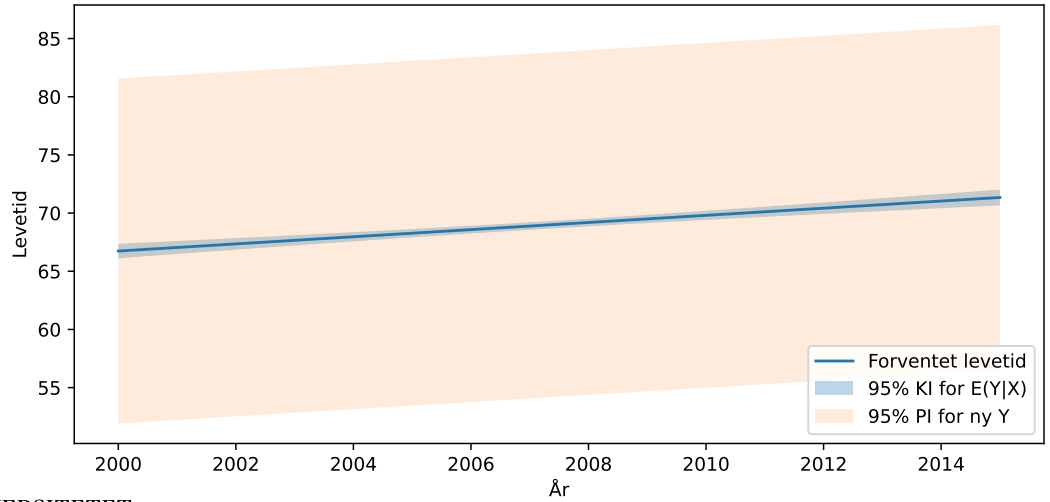
Vi kan visualisere forskjellen på konfidens- og prediksjonsintervallet ved å endre verdiene til en forklaringsvariabel mens de andre holdes fast:

```
# All years
year_grid = np.linspace(LF_X['Year'].min(), LF_X['Year'].max(), 50)
# All other covariates kept at their respective mean value
bmi_mean = LF_X['BMI'].mean()
alcohol_mean = LF_X['Alcohol'].mean()
pop_mean = LF_X['Population'].mean()

grid_df = pd.DataFrame({
    'const': 1.0,
    'Year': year_grid,
    'Alcohol': alcohol_mean,
    'BMI': bmi_mean,
    'Population': pop_mean
})

pred_grid = LF_model.get_prediction(grid_df).summary_frame()
```

Visualisering – konfidens vs prediksjon



Oppsummering

- **Konfidensintervall for** $\mathbb{E}(Y \mid \mathbf{X})$:
  - Beskriver usikkerheten i den sanne regresjonslinjen.
  - Brukes når vi er interessert i **gjennomsnittseffekten**.

- **Prediksjonsintervall for ny** Y:
  - Inkluderer både usikkerhet i linjen og tilfeldige avvik.
  - Brukes når vi vil **predikere en ny observasjon**.

- I `statsmodels`:
  - `get_prediction().summary_frame()` gir begge deler:
    - `mean_ci_*` $\Rightarrow$ konfidensintervall for forventet verdi
    - `obs_ci_*` $\Rightarrow$ prediksjonsintervall for ny observasjon

Kategoriske variabler i regresjon

Mange forklaringsvariabler er kategoriske.

Datasettet om forventet levetid har, for eksempel, en variabel `Status` som tar en av to verdier:

```
print("Status levels:", sorted(LFdata["Status"].unique()))
```

```
Status levels: ['Developed', 'Developing']
```

Vi ønsker å kunne ha med denne type forklaringsvariabler og kombinere dem med numeriske variabler (år, BMI, osv.).

Løsning: Bruk **dummy-variabler** (indikatorvariabler).

Kategoriske variabler i regresjon

Anta at vi kun har `Status` som forklaringsvariabel:

$$y_i = \begin{cases} \alpha_0 + \varepsilon_i & \text{hvis Status = Developed} \\ \alpha_1 + \varepsilon_i & \text{hvis Status = Developing} \end{cases}$$

Da får vi to ulike gjennomsnitt:

- α_0 = forventet levetid for `Developed`
- α_1 = forventet levetid for `Developing`

Dette er formelt en to-utvalgs-situasjon, men vi kan skrive modellen som en vanlig lineær regresjonsmodell ved å lage en dummy-variabel.

Kategoriske variabler i regresjon

Definer en dummy-variabel:

$$x_i = \begin{cases} 1 & \text{hvis Status = Developing} \\ 0 & \text{hvis Status = Developed} \end{cases}$$

Da kan vi skrive modellen som:

$$y_i = \alpha_0 + (\alpha_1 - \alpha_0)x_i + \varepsilon_i.$$

Hvis vi setter $\beta_0 = \alpha_0$ og $\beta_1 = \alpha_1 - \alpha_0$, da får vi en vanlig lineær regresjonsmodell:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Tolkning:

- β_0 = forventet verdi for **referansekategorien** (`Developed`)
- β_1 = **forskjellen** mellom `Developing` og `Developed`

Kategoriske variabler i regresjon

```
LF_X2 = LFdata[["Alcohol", "BMI", "Year", "Population", "Status"]]

# Create dummy-variable for Status (Developing vs Developed)
LF_X2 = pd.get_dummies(LF_X2, columns=["Status"], drop_first=True)
LF_X2['Status_Developing'] = LF_X2['Status_Developing'].astype(int)

# Add a constant
LF_X2 = sm.add_constant(LF_X2)

lmfit2 = sm.OLS(LF_y, LF_X2, missing="drop").fit()
lmfit2.summary2().tables[1].iloc[:,:4].round(3)
```

`drop_first = True` gjør at vi velger en referansekategori (`Developed`)

Kategoriske variabler i regresjon

	Coef.	Std.Err.	t	P>|t|
const	-520.230	73.167	-7.110	0.000
Alcohol	0.158	0.051	3.088	0.002
BMI	0.225	0.009	25.560	0.000
Year	0.292	0.036	8.011	0.000
Population	0.000	0.000	1.637	0.102
Status_Developing	-7.823	0.522	-15.000	0.000

Den nye raden `Status_Developing` i koeffisient-tabellen:

- koeffisienten er estimat av $\alpha_1 - \alpha_0$
- beskriver forskjellen i livslengde mellom `Developing` og `Developed`

Å legge til flere forklaringsvariabler

Når vi legger til `Status` i den opprinnelige modellen (i tillegg til `Alcohol`, `BMI`, `Year`, `Population`):

- Koeffisientene til de andre variablene endres
- Årsak: **korrelasjon** mellom forklaringsvariablene
- Tolkning i multippel regresjon er alltid:
  - «Effekten av denne variabelen **gitt** (betinget på) de andre variablene i modellen»

Katagoriske variabler med flere nivåer

Variablen `Measles` beskriver antall tilfeller per 1000 innbyggere. Også her kan vi endre variablen til å ha katagorier:

```
LFdata["Measles"] = LFdata["Measles"].apply(
    lambda x: "3+" if x > 3 else x
)
```

Vi ønsker modell:

$$y_i = \begin{cases} \alpha_0 + \varepsilon_i & \text{Measles = 0} \\ \alpha_1 + \varepsilon_i & \text{Measles = 1} \\ \alpha_2 + \varepsilon_i & \text{Measles = 2} \\ \alpha_3 + \varepsilon_i & \text{Measles = 3} \\ \alpha_4 + \varepsilon_i & \text{Measles = 3+} \end{cases}$$

Vi har fem nivåer $\Rightarrow$ fem forskjellige gjennomsnitt

Katagoriske variabler med flere nivåer

Vi trenger nå flere dummy-variabler. For eksempel:

$x_{i,1} = I(\text{Measles} = 1)$, $x_{i,2} = I(\text{Measles} = 2)$, $x_{i,3} = I(\text{Measles} = 3)$, $x_{i,4} = I(\text{Measles} = 3+)$

der $I(\cdot)$ er 1 hvis utsagnet er sant, ellers 0.

Nivået `Measles = 0` er referanse hvor $x_{i,1} = x_{i,2} = x_{i,3} = x_{i,4} = 0$.

Modellen blir:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \varepsilon_i.$$

med

$$\beta_0 = \alpha_0, \quad \beta_1 = \alpha_1 - \alpha_0, \ldots$$

Katagoriske variabler med flere nivåer

```
LF_X3 = LFdata[["Alcohol", "BMI", "Year", "Population", "Status", "Measles"]]

# Dummy-variables for Status and Measles
LF_X3 = pd.get_dummies(LF_X3, columns=["Status"], drop_first=True)
LF_X3 = pd.get_dummies(LF_X3, columns=["Measles"], drop_first=True)

# Convert bool to int (0/1)
bool_cols = LF_X3.select_dtypes(include="bool").columns
LF_X3[bool_cols] = LF_X3[bool_cols].astype(int)

# Add a constant
LF_X3 = sm.add_constant(LF_X3)

lmfit3 = sm.OLS(LF_y, LF_X3, missing="drop").fit()
lmfit3.summary2().tables[1].iloc[:,:4].round(3)
```

Katagoriske variabler med flere nivåer

	Coef.	Std.Err.	t	P>|t|
const	-488.775	72.153	-6.774	0.000
Alcohol	0.132	0.051	2.620	0.009
BMI	0.209	0.009	23.679	0.000
Year	0.278	0.036	7.731	0.000
Population	0.000	0.000	2.436	0.015
Status_Developing	-8.387	0.518	-16.192	0.000
Measles_1	-2.869	0.808	-3.550	0.000
Measles_2	-1.059	1.002	-1.057	0.291
Measles_3	-0.675	1.227	-0.550	0.582
Measles_3+	-2.941	0.349	-8.437	0.000

- Vi får nå flere rader, `Measles_1`, `Measles_2`, ...
- Hver koeffisient beskriver **forskjellen** i forhold til referansen `Measles = 0`

Oppsummering – kategoriske variabler

- **Idéen**:
  - Kategorier $\Rightarrow$ dummy-variabler (0/1)
  - Én kategori velges som referanse, får ikke egen kolonne
- **To nivåer**:
  - én dummy, koeffisienten = nivå-forskjell
- **Flere nivåer**:
  - k nivåer $\Rightarrow$ $k - 1$ dummy-variabler
  - hver koeffisient = forskjell mot referansekategorien
- I multippel regresjon:
  - Husk å tolke alle effekter betinget på de andre variablene i modellen.