

- **Det gjøres opptak av forelesningen**
  - For de som har annen undervisning som kolliderer, og til eksamensforberedelse.
- Vil dere stille spørsmål, men ikke vil bli en del av opptaket, bruk "chat" funksjonen.
- Opptaket vil bli lagret på emnesiden til PSY2014 UiO, og en lenke vil tilgjengelig for de som følger kurset.
- Opptaket skal bli slettet etter 2022.



# PSY2014 - KVANTITATIV METODE

Forelesning 5: Kategoriske prediktorer, interaksjon, mediering  
Nikolai Czajkowski

1. Avsluttende om F-test og konfidensintervaller
2. Kollinearitet
3. Kategoriske uavhengige variabler.
4. Modellering av interaksjon
5. Mediering
6. Modellbygging / Modellvalg

## Hypotesene

$$H_0 : b_j = 0$$

$$H_1 : b_j \neq 0$$

## Test statistikken

$$t = \frac{\hat{b}_j}{\hat{SE}(\hat{b}_j)}$$

## Samplingfordelingen under $H_0$ :

Under  $H_0$  følger t-statistikken en  $t(n - p - 1)$  fordeling.



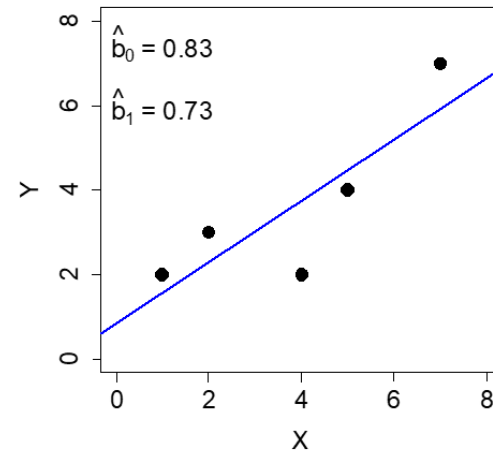
Konfidensintervaller (CIs) er en annen måte å uttrykke usikkerhet til estimatene.

$$\hat{b}_1 = 10 (8, 12)$$

- Enhver verdi i et 95% konfidensintervall utgjør en hypotese om parameterets verdi som du *ikke* kan forkaste på et 0.05 nivå.
  - Jeg kan ikke forhaste hypotesen om at  $b_1$  er 9 i populasjonen.
- Enhver verdi *utenfor* intervallet kan forkastes på et 0.05 nivå.
  - Jeg kan forhaste hypotesen om at  $b_1$  er 13 i populasjonen.
- Et konfidensintervall gir mer informasjon enn en p-Verdi.

# UTREGNING AV KONFIDENSINTERVALLER

```
> confint(M1)
                2.5 %    97.5 %
(Intercept) -2.9597439  4.626411
X            -0.1421214  1.598262
```



## 2. For hånd

95% Konfidensintervall :  $\hat{b} \pm t_{df, \alpha} \times SE(\hat{b})$

↑ Dette gir den kritiske verdien ved en t-fordeling med df-frihetsgrader og et gitt alfa nivå.

df	Level of Significance for One-Tailed					
	0.25	0.20	0.15	0.10	0.05	0.025
	Level of Significance for Two-Tailed					
	0.50	0.40	0.30	0.20	0.10	0.05
1	1.000	1.376	1.963	3.078	6.314	12.706
2	0.816	1.061	1.386	1.886	2.920	4.303
3	0.765	0.978	1.250	1.638	2.353	3.182
4	0.741	0.941	1.190	1.533	2.132	2.776

Øvre grense:  $\hat{b} + t_{n-p-1} \times SE(\hat{b}) = 0.728 + 3.18 \times 0.273 = 1.598$

Nedre grense:  $\hat{b} - t_{n-p-1} \times SE(\hat{b}) = 0.728 - 3.18 \times 0.273 = -0.142$

## F-TESTEN: TESTING AV FLERE REGRESJONSKOEFFSIENTER SAMLET



## Hypotesene

$$H_0 : b_1 = b_2 = b_3 = \dots = b_p = 0$$

$$H_1 : \text{minst en } b_j \neq 0$$

## Test statistikken

$$F = \frac{MSM}{MSE}$$

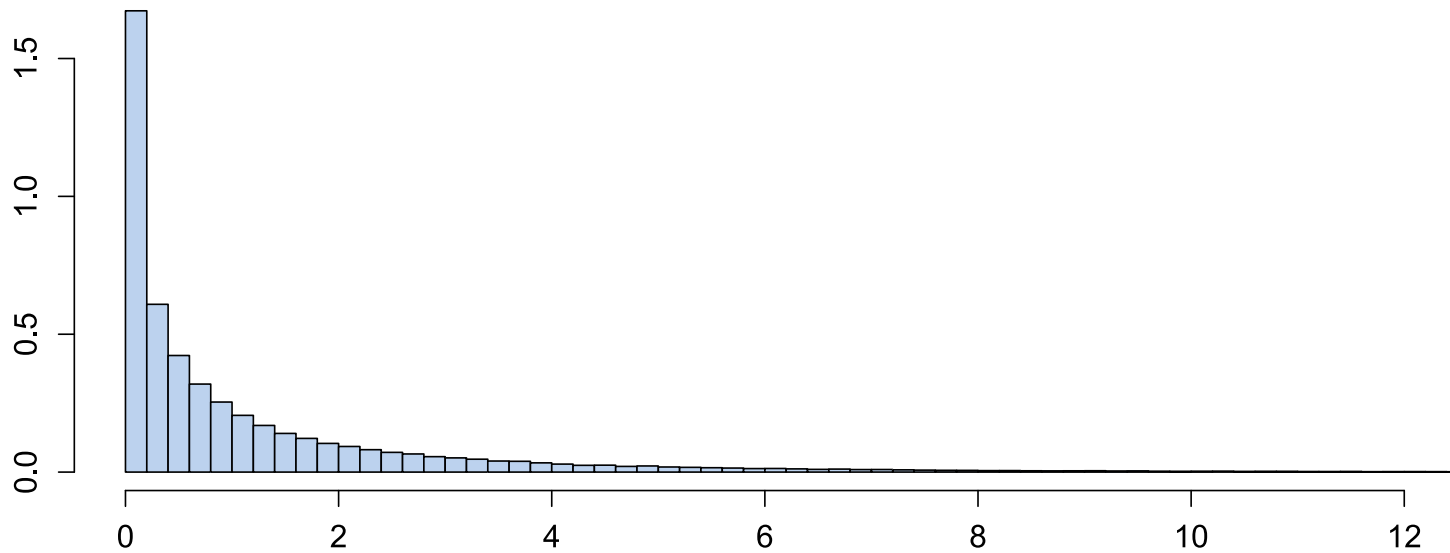
$$MSM = \frac{SSM}{df_{SSM}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{p}$$

$$MSE = \frac{SSE}{df_{SSE}} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - p - 1}$$

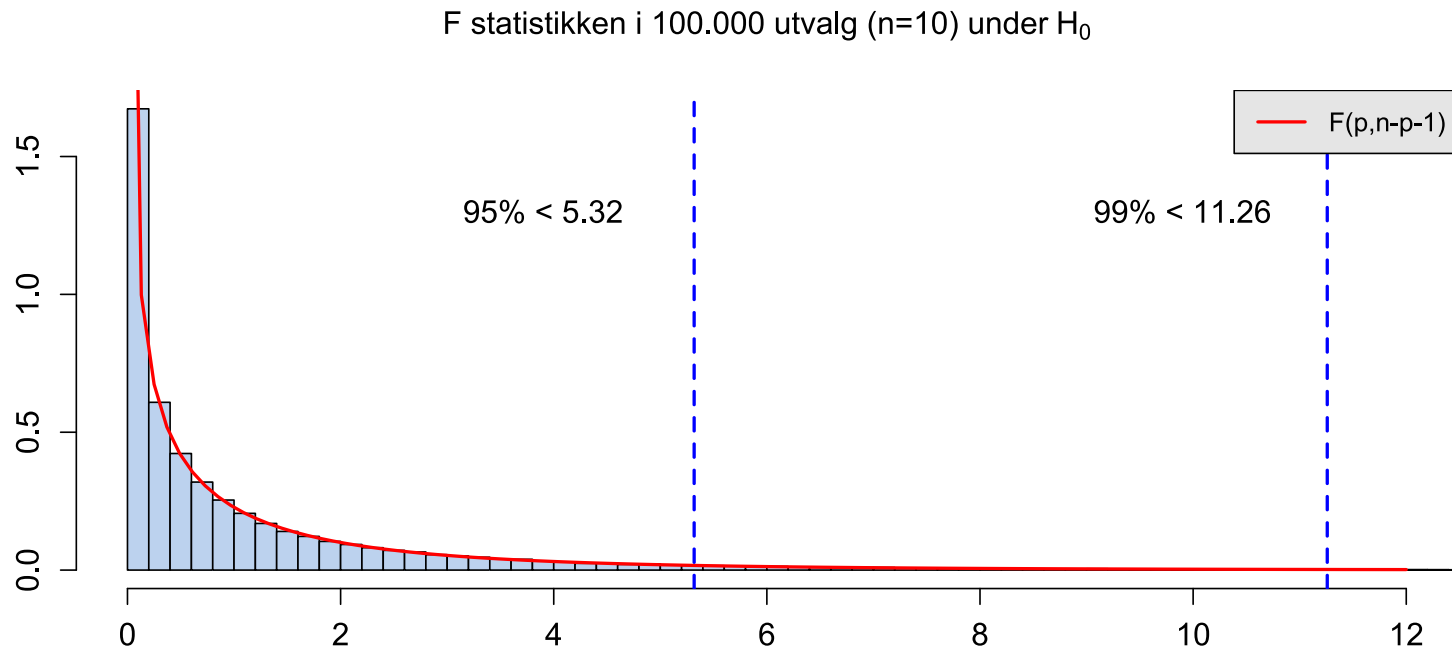
## Samplingfordelingen under $H_0$

Under  $H_0$  følger F-statistikken en  $F(p, n - p - 1)$  fordeling.

F statistikken i 100.000 utvalg ( $n=10$ ) under  $H_0$

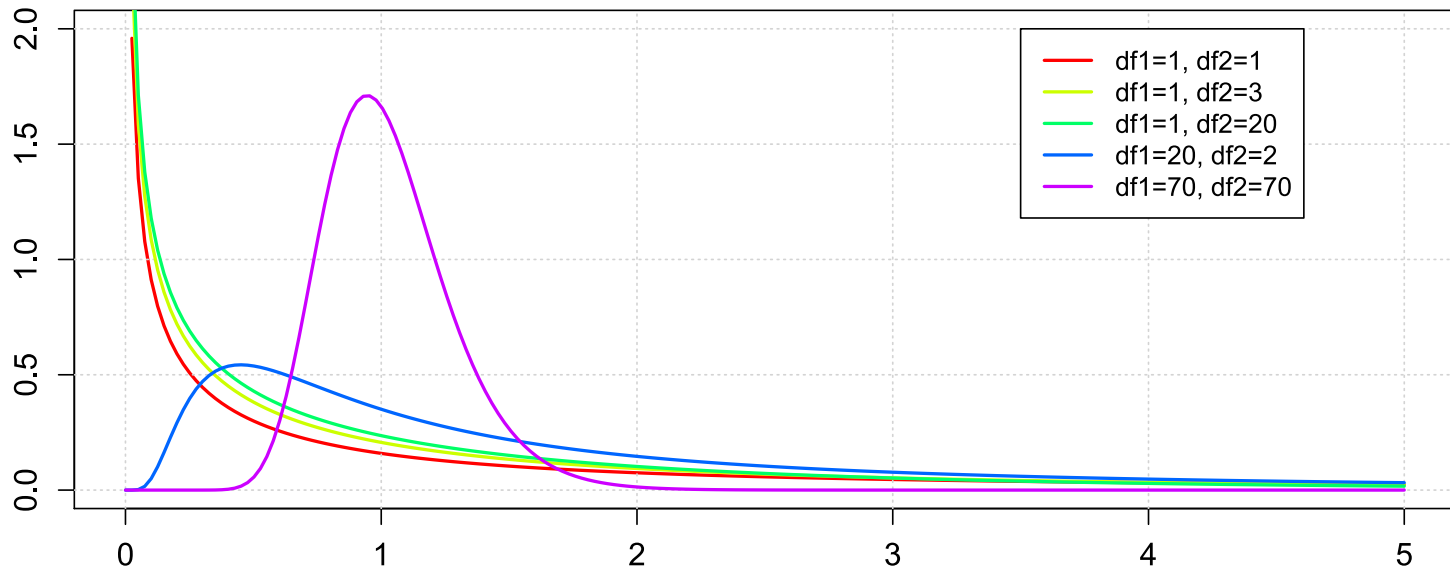


# SAMPLINGFORDELINGEN TIL F (UNDER $H_0$ )



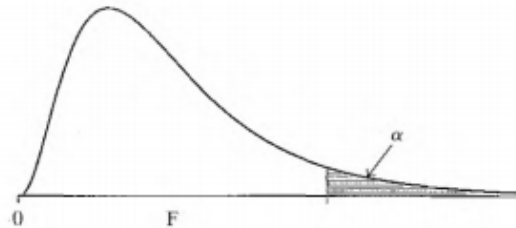
For the F-distribution, only a one tailed test is reasonable.

# F-FORDELINGEN



# F-TABELLEN (1)

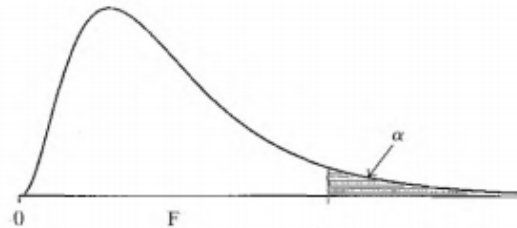
TABLE D: F Distribution



$\alpha = .05$										
$df_2$	$df_1$									
	1	2	3	4	5	6	8	12	24	$\infty$
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07

# F-TABELLEN (2)

TABLE D: F Distribution



$\alpha = .05$										
$df_2$	$df_1$									
	1	2	3	4	5	6	8	12	24	$\infty$
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07

# UTSKRIFT AV F-STATISTIKKEM

```
lm(formula = Y ~ X)

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.8333      1.1919   0.699   0.5348
X            0.7281      0.2734   2.663   0.0762 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.306 on 3 degrees of freedom
Multiple R-squared:  0.7027,    Adjusted R-squared:  0.6036
F-statistic:  7.09 on 1 and 3 DF,  p-value: 0.07616
```

```
> anova(M1)
Analysis of Variance Table

Response: Y
      Df Sum Sq Mean Sq F value    Pr(>F)
X       1 12.086 12.0860   7.0899 0.07616 .
Residuals 3  5.114  1.7047
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





# UNDERLIGE RESULTATER

```
summary(lm(Y~X1))
```

```
## (Intercept)  -1.2163      0.5793  -2.100   0.0401 *
## X1           2.8746      0.5710   5.035  4.95e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.384 on 58 degrees of freedom
## Multiple R-squared:  0.3041,    Adjusted R-squared:  0.2921
## F-statistic: 25.35 on 1 and 58 DF,  p-value: 4.954e-06
```

```
summary(lm(Y~X2))
```

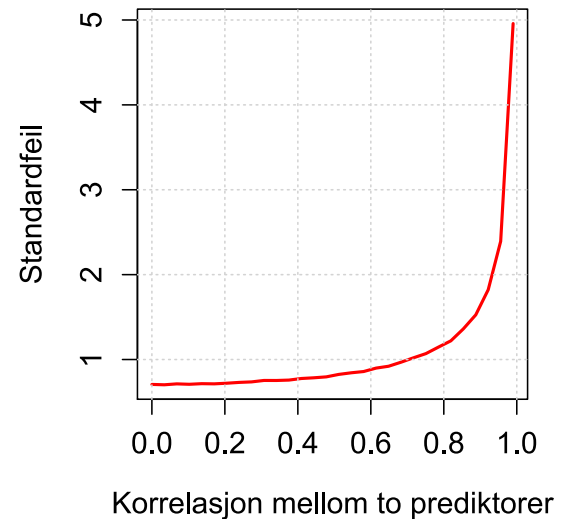
```
## (Intercept)  -1.0249      0.5585  -1.835   0.0716 .
## X2           2.9151      0.5379   5.419  1.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.282 on 58 degrees of freedom
## Multiple R-squared:  0.3361,    Adjusted R-squared:  0.3247
## F-statistic: 29.37 on 1 and 58 DF,  p-value: 1.208e-06
```

```
summary(lm(Y~X1+X2))
```

```
## (Intercept)  -1.1017      0.5719  -1.926   0.0590 .
## X1           0.8634      1.2485   0.692   0.4921
## X2           2.1708      1.2043   1.803   0.0768 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.302 on 57 degrees of freedom
## Multiple R-squared:  0.3417,    Adjusted R-squared:  0.3186
## F-statistic: 14.79 on 2 and 57 DF,  p-value: 6.699e-06
```

**Kollinearitet** er graden av lineær sammenheng mellom flere forklaringsvariabler i en multippel regresjonsmodell.

Blir denne for høy (korrelasjon  $\gg 0.8$ ) kan det medføre problemer.



## Konsekvenser av kollinearitet

- b'ene er ikke til å stole på grunnet stør økning i standardfeilene.
- Vanskelig å vurdere hvilken av de uavhengige variablene er av betydning.
- Justert R<sup>2</sup> kan begynne å synke.
- Tilpasning av modellen er ikke påvirket.

```
library(car)
m1<-lm(Inntekt2~Alder2+Erfaring2)
cor(Alder2, Erfaring2)
```

```
## [1] 0.7579404
```

```
vif(m1)
```

```
##      Alder2  Erfaring2
## 2.350031  2.350031
```

- Se etter store korrelasjoner mellom de uavhengige variablene ( $>0.8$ ).
- VIF (variance inflation factor) er en diagnostisk statistikk som angir grad av multikollinearitet.
  - VIF tallfester hvor mye en uavhengig variabel øke standardfeilen
  - Høye VIF-verdier (over 4 eller 5) tyder på kollinearitetsproblemer.



## GIR DETTE MENING?

Anta at vi ønsker å se på om folk i tre forskjellige byer er lykkeligere. Anta at vi hadde variabelen `by` kodet slik:

Hvordan skulle vi forstå følgende resultater (tolkningen av  $b_0$  og  $b_1$ )?

by	verdi
Oslo	1
Stavanger	2
Trondheim	3

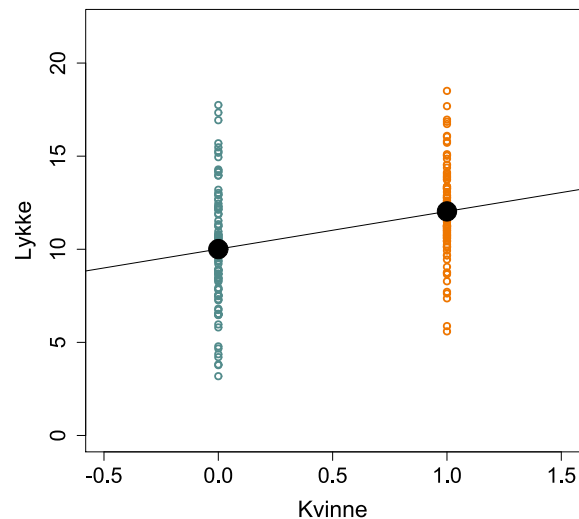
```
summary(lm(lykke~by, data=hap))
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.2222     0.2631 133.882   <2e-16 ***
## by           0.4930     0.1974   2.498    0.0128 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En **dummy variabel** (indikator variabel) tar bare verdiene 0 eller 1.

Var	MANN
Kvinner	0
Menn	1

$$LYKKE_i = b_0 + b_1 \cdot MANN_i$$



Det er kanskje ikke meningsfullt å snakke om “én enhets økning i MANN”, men siden variabelen bare har to nivåer kan vi si at  $b_1$  er forventet forskjell i lykke mellom men og kvinner.

## KATEGORISKE UAVHENGIGE VARIABLER (FAKTORER)

Original Koding	Dummy <sub>1</sub>	Dummy <sub>2</sub>
1 (Oslo)	0	0
2 (Stavanger)	1	0
3 (Trondheim)	0	1

$$Lykke = b_0 + b_1 \cdot Dummy_1 + b_2 \cdot Dummy_2$$

- By er en kategorisk variabel med flere nivåer, og det gir ikke mening å snakke om betydningen av en økning på en enhet.
- For å inkludere kategoriske variabler som uavhengige variabler i regresjonsmodeller kan vi rekode dem til dummy variabler.
- I kodeingen under blir Oslo *referansekategori*.

```
M2 <- lm(formula = Lykke ~ D1 + D2, data = dt)
summary(M2)
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.8120     0.2181  44.983 < 2e-16 ***
## D1           -0.5694     0.3085  -1.846  0.0654 .
## D2            1.4723     0.3085   4.773 2.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.085 on 597 degrees of freedom
## Multiple R-squared:  0.07249,    Adjusted R-squared:  0.06939
## F-statistic: 23.33 on 2 and 597 DF,  p-value: 1.754e-10
```

Original Koding	Dummy1	Dummy2
1 (Oslo)	0	0
2 (Stavanger)	1	0
3 (Trondheim)	0	1

$$Lykke_O = 9.81 + (-0.57) \cdot 0 + 1.47 \cdot 0 = 9.81$$

$$Lykke_S = 9.81 + (-0.57) \cdot 1 + 1.47 \cdot 0 = 9.24$$

$$Lykke_T = 9.81 + (-0.57) \cdot 0 + 1.47 \cdot 1 = 11.28$$



# T-TEST

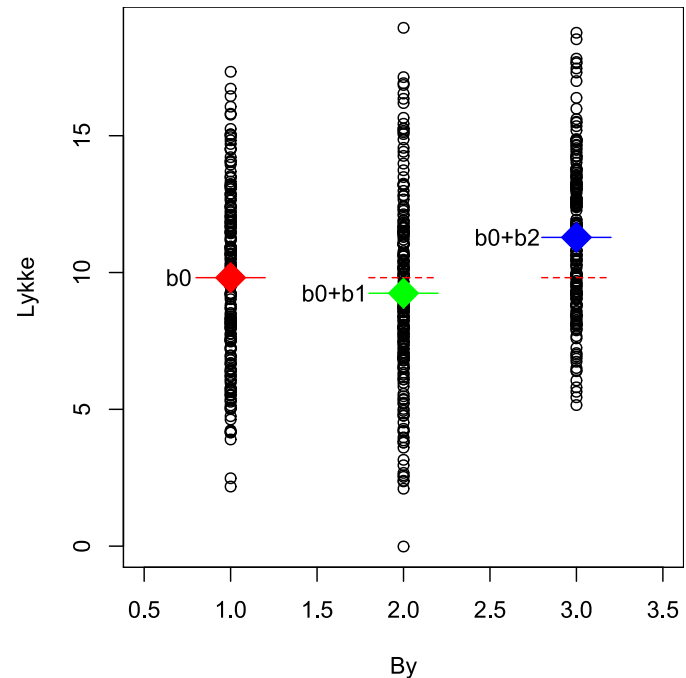
```
M2 <- lm(formula = Lykke ~ D1 + D2, data = dt)
summary(M2)
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.8120     0.2181  44.983 < 2e-16 ***
## D1           -0.5694     0.3085  -1.846  0.0654 .
## D2            1.4723     0.3085   4.773 2.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.085 on 597 degrees of freedom
## Multiple R-squared:  0.07249,    Adjusted R-squared:  0.06939
## F-statistic: 23.33 on 2 and 597 DF,  p-value: 1.754e-10
```

Signifikansen av forskjell i gjennomsnitt mellom referansegruppen og hver av de andre kan vurderes med *t-test*.

$$H_0 : b_2 = 0 \quad (\mu_{oslo} = \mu_{stavanger})$$

$$H_0 : b_3 = 0 \quad (\mu_{oslo} = \mu_{trondheim})$$



Original Koding	Dummy1	Dummy2
1 (Oslo)	0	0
2 (Stavanger)	1	0
3 (Trondheim)	0	1

# F-TEST

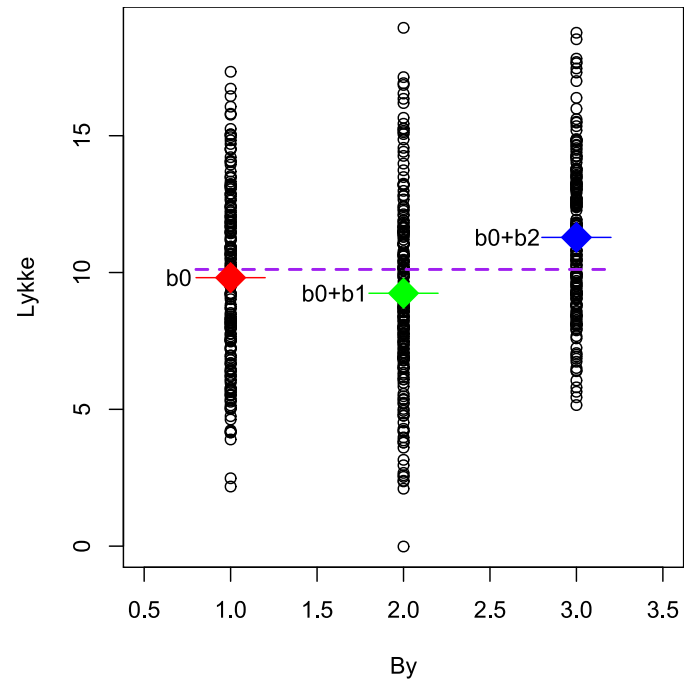
```
M2 <- lm(formula = Lykke ~ D1 + D2, data = dt)
summary(M2)
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.8120     0.2181  44.983 < 2e-16 ***
## D1            -0.5694     0.3085  -1.846  0.0654 .
## D2             1.4723     0.3085   4.773 2.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.085 on 597 degrees of freedom
## Multiple R-squared:  0.07249,    Adjusted R-squared:  0.06939
## F-statistic: 23.33 on 2 and 597 DF,  p-value: 1.754e-10
```

Hypotesen om at det er samme gjennomsnitt i *alle* gruppen kan vurderes med en F-test.

$$H_0 : b_2 = b_3 = 0$$

$$(\mu_{oslo} = \mu_{stavanger} = \mu_{trondheim})$$



Original Koding	Dummy1	Dummy2
1 (Oslo)	0	0
2 (Stavanger)	1	0
3 (Trondheim)	0	1

# BETYDNINGEN AV REFERANSEKATEGORI

Original Koding	Dummy <sub>1</sub>	Dummy <sub>2</sub>	Dummy <sub>3</sub>	Dummy <sub>4</sub>
1 (Oslo)	0	0	1	0
2 (Stavanger)	1	0	0	1
3 (Trondheim)	0	1	0	0

```
M2 <- lm(formula = Lykke ~ D1 + D2, data = dt)
summary(M2)
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.8120      0.2181  44.983 < 2e-16 ***
## D1            -0.5694      0.3085  -1.846  0.0654 .
## D2             1.4723      0.3085   4.773 2.29e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.085 on 597 degrees of freedom
## Multiple R-squared:  0.07249,    Adjusted R-squared:  0.06939
## F-statistic: 23.33 on 2 and 597 DF,  p-value: 1.754e-10
```

```
M3 <- lm(formula = Lykke ~ D3 + D4, data = dt)
summary(M3)
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.2843      0.2181  51.733 < 2e-16 ***
## D3           -1.4723      0.3085  -4.773 2.29e-06 ***
## D4           -2.0417      0.3085  -6.619 8.08e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.085 on 597 degrees of freedom
## Multiple R-squared:  0.07249,    Adjusted R-squared:  0.06939
## F-statistic: 23.33 on 2 and 597 DF,  p-value: 1.754e-10
```

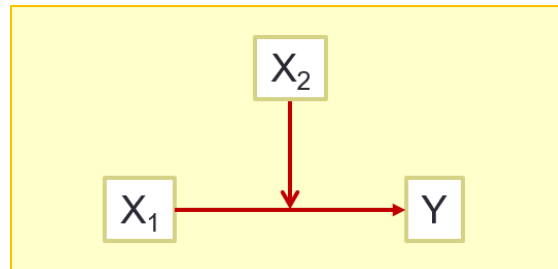


## **Hva er forholdet mellom arbeidserfaring, kjønn og inntekt?**

- Har folk med lengre erfaring høyere lønn?
  - *Er det en effekt av erfaring?*
- Tjener kvinner mer eller mindre enn menn?
  - *Er det en effekt av kjønn?*
- Er bidraget av arbeidserfaring på inntekt det samme hos menn og kvinner?
  - *Er det en interaksjon mellom kjønn og erfaring?*

## INTERAKSJON (MODERASJON) (A 11.5)

- Effekter er **additive** dersom den kombinerte effekten av et lik summen av enkelteffektene.
  - Universitetsgrad gir 1000K NOK forventet økt lønn og 3 år arbeidserfaring gir 50K NOK. Effekten er additiv om forventet lønn for en med både universitetsgrad og 3 år arbeidserfaring er 150K NOK.
  - Genvariant X bidrar til 0.5 cm høyde, genvariant Y bidrar til 0.3 cm høyde, effekten av å ha begge variantene er 0.8 cm høyde.



- **Interaksjon** i statistikk er når effekten av to variabler *ikke er additiv*, men når styrken på forholdet mellom to variabler avhenger av en tredje variabel.
  - Genvariant X eller Y resulterer ikke i sykdom alene. Sykdom oppstår kun når du arver begge variantene.

Inntekt (100)	Mann	Erfaring	Mann * Erfaring
300	0	10	0
320	1	10	10
410	0	20	0
370	1	15	15
550	0	10	0
430	1	5	5

- Over er kvinner kodet 0, mens menn kodet 1.
- For å modellere interaksjon vil jeg definere et *kryssprodktledd* som er gitt av produktet mellom verdiene på de to uavhengige variablene.

Interaksjon modelleres vanligvis med et kryssproduktledd.

I: Inntekt (1000 NOK)

M: Mann (Kvinner=0, Menn=1)

E: Erfaring (år)

$$I_i = b_0 + b_1 \cdot M_i + b_2 \cdot E_i + b_3 \cdot \underbrace{M_i \cdot E_i}_{\text{kryssprodukt}}$$

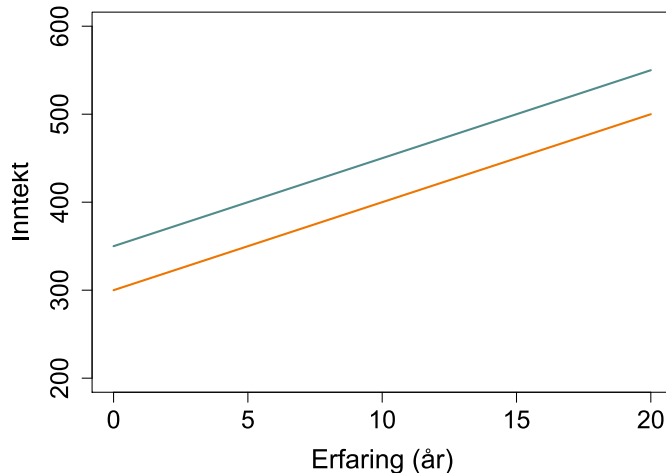
$$\begin{aligned}\text{Kvinner: } I_k &= b_0 + b_1 \cdot 0 + b_2 \cdot E + b_3 \cdot 0 \cdot E \\ I_k &= b_0 + b_2 \cdot E\end{aligned}$$

$$\begin{aligned}\text{Menn: } I_m &= b_0 + b_1 \cdot 1 + b_2 \cdot E + b_3 \cdot 1 \cdot E \\ I_m &= b_0 + b_1 + b_2 \cdot E + b_3 \cdot E \\ I_m &= \underbrace{(b_0 + b_1)}_{\text{ny intercept}} + \underbrace{(b_2 + b_3)}_{\text{nytt stigningstall}} \cdot E\end{aligned}$$



Kvinner:  $I_k = b_0 + b_2 \cdot E$

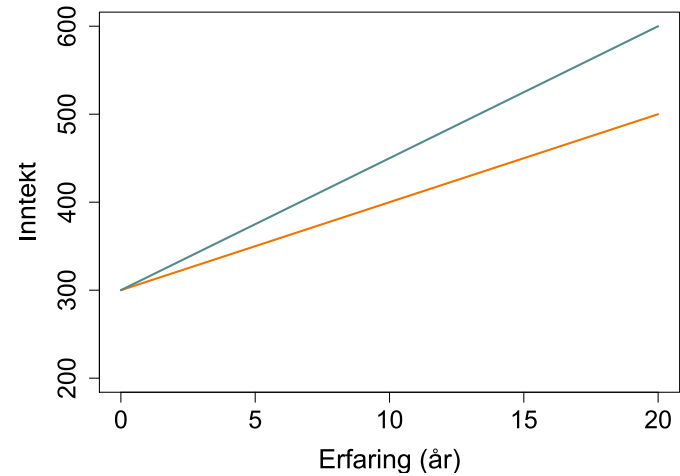
Menn:  $I_m = (b_0 + b_1) + (b_2 + b_3) \cdot E$



$$b_0 = 300, b_1 = 50, b_2 = 10, b_3 = 0$$

$$I_k = 300 + 10 \cdot E$$

$$I_m = (300 + 50) + (10 + 0) \cdot E$$



$$b_0 = 300, b_1 = 0, b_2 = 10, b_3 = 5$$

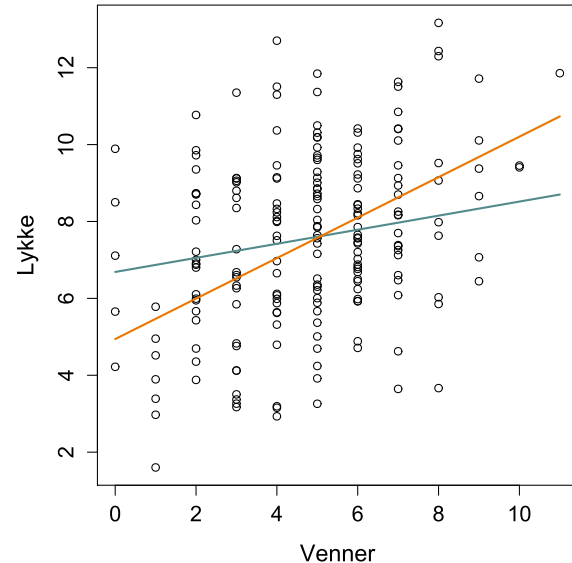
$$I_k = 300 + 10 \cdot E$$

$$I_m = (300 + 0) + (10 + 5) \cdot E$$

# MODELLERING AV INTERAKSJON I R

```
M4 <- lm(Lykke~Venner+Mann+Venner:Mann, data=dt)
summary(M4)
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.94162    0.52011   9.501  < 2e-16 ***
## Venner       0.52648    0.09633   5.465 1.39e-07 ***
## Mann        1.74562    0.72678   2.402  0.0172  *
## Venner:Mann -0.34342    0.13842  -2.481  0.0139  *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.12 on 196 degrees of freedom
## Multiple R-squared:  0.1451,    Adjusted R-squared:  0.132
## F-statistic: 11.09 on 3 and 196 DF,  p-value: 9.373e-07
```



$$Lykke = b_0 + b_1 \cdot Venner_i + b_2 \cdot Mann_i + b_3 \cdot Venner_i \cdot Mann_i$$

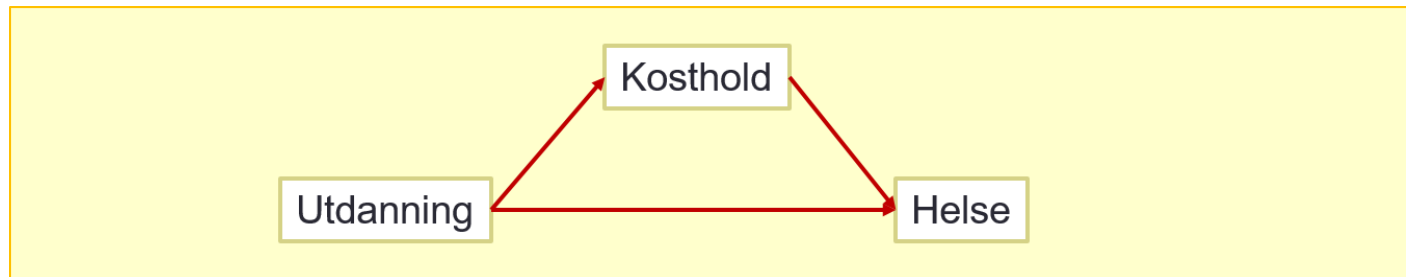
$$Lykke_k = 4.94 + 0.53 \cdot Venner_i$$

$$Lykke_m = 4.94 + 0.53 \cdot Venner_i + 1.75 \cdot 1 + (-0.34) \cdot Venner_i \cdot 1$$

$$Lykke_m = (4.94 + 1.75) + (0.53 - 0.34) \cdot Venner_i$$



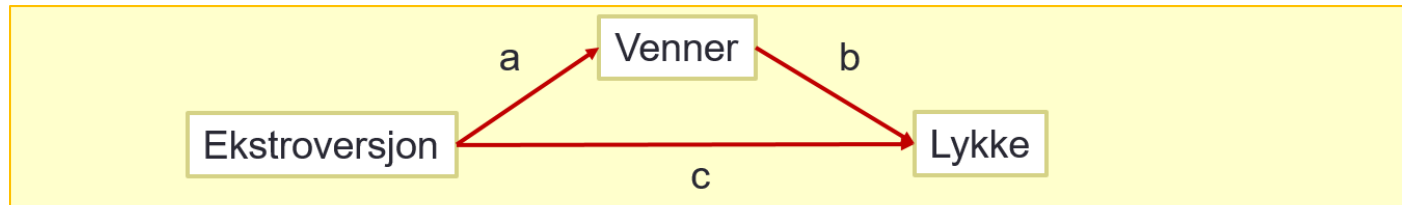
*Hvordan påvirker utdanning helse? Trolig er det ikke en direkte helsegevinst av å gå på skolen, men effekten er mediert gjennom andre variabler.*



**Mediator** er en variabel som formidler deler av effekten/risikoen av andre uavhengige variabler på utfallet.

- På rent statistisk grunnlag kan du ikke skille mellom en mediator og en konfunderende variabel. Dette må gjøres på teoretisk grunnlag.

# KLASSISK MÅTE Å PÅVISE MEDIATOR



- **Trinn 1:** Vis at den første variabelen (ekstroversjon) er assosiert med den avhengige [sti c].
- **Trinn 2:** Vis at den første variabelen er assosiert med mediatoren [sti a].
- **Trinn 3:** Vis at mediatoren er assosiert med utfallet [sti b].
- **Trinn 4:** For å kunne hevde at effekten av ekstroversjon på lykke medieres av antall venner, bør effekten av ekstroversjon blir vesentlig redusert når venner også inkluderes i modellen.

```
summary(lm(lykke~ekstroversjon, data=hap))
```

```
## (Intercept)  31.25640    0.53182  58.772  <2e-16 ***  
## ekstroversjon  0.44187    0.05017   8.807  <2e-16 ***
```

```
summary(lm(venner~ekstroversjon, data=hap))
```

```
## (Intercept)   2.26110    0.20388   11.09  <2e-16 ***  
## ekstroversjon 0.28448    0.01923   14.79  <2e-16 ***
```

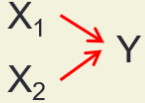
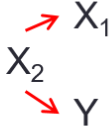
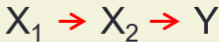
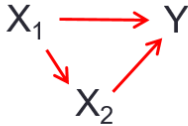
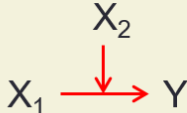
```
summary(lm(lykke~venner, data=hap))
```

```
## (Intercept) 31.30758    0.49011  63.879  <2e-16 ***  
## venner      0.86047    0.09041   9.518  <2e-16 ***
```

```
summary(lm(lykke~ekstroversjon+venner, data=hap))
```

```
## (Intercept)  29.88385    0.56840  52.575  < 2e-16 ***  
## ekstroversjon 0.26918    0.05707   4.717 2.99e-06 ***  
## venner        0.60703    0.10383   5.847 8.27e-09 ***
```

# OPPSUMMERING AV 3-VARIABEL ASSOSIASJONER

Graph	Name of relationship	What happens after controlling for $X_2$
 <pre> graph LR     X1 --&gt; Y     X2 --&gt; Y         </pre>	Multiple causes	If $X_1$ and $X_2$ are uncorrelated, the association between $X_1$ and $Y$ does not change
 <pre> graph LR     X2 --&gt; X1     X2 --&gt; Y         </pre>	Spurious $X_1$ $Y$ association	Association between $X_1$ and $Y$ disappears
 <pre> graph LR     X1 --&gt; X2     X2 --&gt; Y         </pre>	Chain relationship (complete mediation)	Association between $X_1$ and $Y$ disappears
 <pre> graph LR     X1 --&gt; Y     X1 --&gt; X2     X2 --&gt; Y         </pre>	Both direct and indirect effects of $X_1$ on $Y$ (Partial mediation)	Association between $X_1$ and $Y$ changes but does not disappear
 <pre> graph LR     X1 --&gt; Y     X2 --&gt;  moderates  X1_Y_path         </pre>	Interaction	Association between $X_1$ and $Y$ varies according to level of $X_2$



# HVORDAN VELGE ENDLIG MODELL?

Hvorvidt en uavhengige variabel vil være signifikant, avhenger også av de andre uavhengige variablene i modellen (dersom de er korrelerte). Hvordan skal du gå frem for å bestemme hvilke uavhengige variabler som skal inngå i den endelige modellen?

## 1. Hierarkisk regresjon

- Samle uavhengige variabler i grupper (blokker), og sjekk om tilpasningen blir vesentlig (signifikant) bedre når du legger til en blokk.

## 2. Skrittvis (stepwise)

- En helt automatisert fremgangsmåte å velge prediktorer.

## 3. Informasjonskriterier

- Bruk statistikker som finner den beste balansen mellom *undertilpasning* og *overtilpasning*.



# METODE 1: HIERARKISK REGRESJON

- I hierarkisk regresjon blir grupper av uavhengige variabler trinnvis lagt til.
- F-testen kan benyttes til å vurdere om et trinn signifikant forbedrer modellen.

```
m4a<-lm(lykke~alder+kjønn,data=hap)
m4b<-lm(lykke~alder+kjønn+inntekt+utdanning,data=hap)
anova(m4a, m4b)
```

```
## Analysis of Variance Table
##
## Model 1: lykke ~ alder + kjønn
## Model 2: lykke ~ alder + kjønn + inntekt + utdanning
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      597 9708.4
## 2      595 7754.5  2      1954 74.964 < 2.2e-16 ***
## ---
```

$$F = \frac{(SSE_R - SSE_C)/df_1}{SSE_C/df_2}$$

$$F = \frac{(9708.4 - 7754.5)/2}{(7754.5/597)}$$

$$F \approx 74.964$$

## METODE 2: SKRITTVIS (STEPWISE)

- **Skrittvis (stepwise)** regresjon er en algoritme som automatisk velger ut variabler for en regresjonsmodell.
  - **Forward selection:** Start uten uavhengige variabler i modellen. Test hvor mye hver enkelt nye variabel ville forbedret modellen, og legg til den som forbedrer modellen mest. Gjenta denne prosessen til ingen nye variabler signifikant forbedrer modellen.

```
library(olsrr)
model <- lm(lykke~alder+kjønn+inntekt+utdanning+ekstroversjon+søsken+hjem, data = hap)
ols_step_forward_p(model)
```

```
##
##                               Selection Summary
## -----
##      Variable                Adj.
## Step      Entered      R-Square  R-Square    C(p)      AIC      RMSE
## -----
##    1   utdanning      0.1150    0.1136   175.8765   3318.2936   3.8304
##    2   ekstroversjon  0.2121    0.2094    93.2391   3250.6083   3.6173
##    3   inntekt        0.3005    0.2969    18.1470   3181.2188   3.4113
##    4   kjønn         0.3176    0.3130     5.2229   3168.3605   3.3721
##    5   søsken         0.3209    0.3151     4.3618   3167.4695   3.3669
## -----
```

- Stepwise regresjon er en fulstendig automatisert prosess der dataene får drive modellen. Mange misliker at statistiske kriterier får informere teori uten ukritisk vurdering eller teoretisk forankring (det er ingen garanti for at den endelige modellen "gir mening").
- Det er ingen garanti for at fremover og bakover-prosedylene vil resultere i at du velger ut det *samme* settet med uavhengige variabler.
- Du bruker det samme datasettet både til å velge prediktorer, og vurdere signifikans.

# OVER VS. UNDERTILPASNING (SCYLLA OG CHARYBDIS)



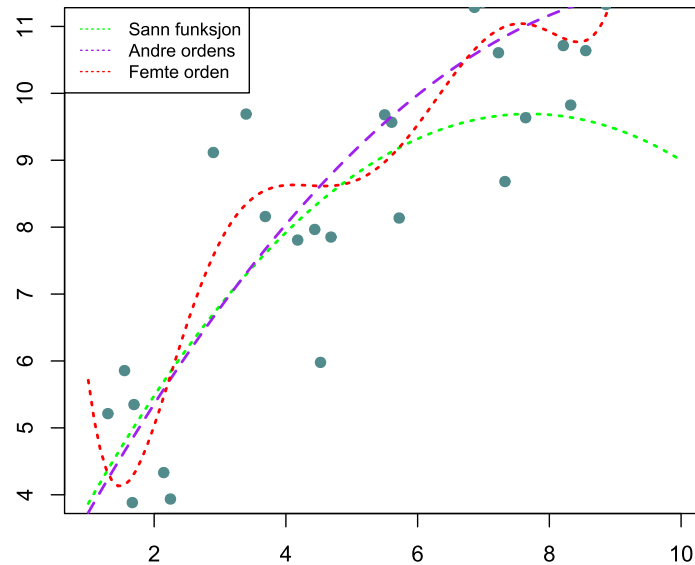
**Overtilpasning:** Dårlig prediksjon som følge av at du lærer for mye fra data.

**Undertilpasning:** Dårlig prediksjon som følge av at du lærer for lite fra data.

- Du kan alltid på bedre tilpasning ved å inkludere flere prediktorer.
- Alle datasett inneholder både systematiske og usystematiske effekter (støy), så komplekse modeller kan predikere *mindre* varians i nye dataset.

<https://xcelab.net/rm/statistical-rethinking/>

# HVA SKJER NÅR MODELLEN BLIR FOR KOMPLEKS?

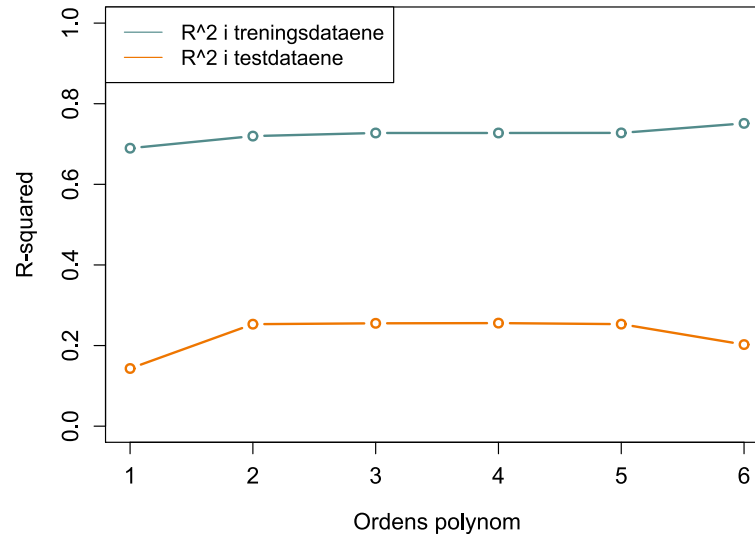


Over: Den grønne kurven er "sann". Den røde passer best, men tilpasser seg mye av støyen i dataene.



- **Kryssvalidering** innebærer å dele datasettet ditt inn i et *treningssett* og et *testsett*. Så bygger du modellen utifra treningssettet, og gjør den endelige vurderingen av modellen etter hvor godt den presterer i testsettet.
- Kryssvalidering er en hjørnesten i moderne maskinlæringsmetoder.

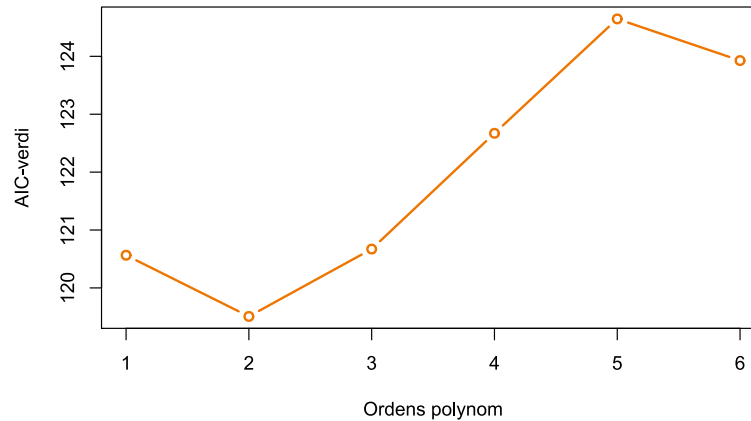
## K-FOLD KRYSSVALIDERING (2)



Merk:

1. Det er lett å forklare varians, det er mye vanskeligere å predikere nye data.
2. Først øker forklart varians i testutvalget, men ettersom modellen blir mer og mer kompleks, begynner ytelsen i tesdataene å *synke*.

## METODE 3: INFORMASJONSKRITERIER (AIC / BIC)



- **Informasjonskriterier** er en klasse statistikker som er utviklet for å optimalt balansere mellom over og undertilpasning, tilpasning og kompleksitet.
- Modellen som velges er den som har *lavest* verdi på informasjonskriteriet, her Akaike information criterion (AIC).