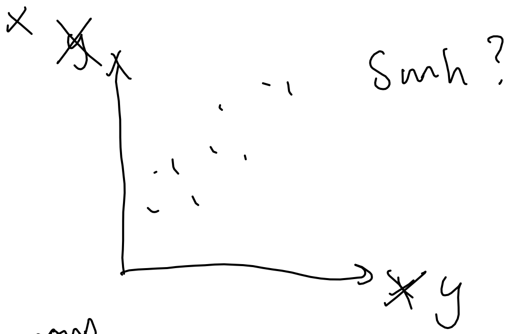
Regresjonsanalyse: Konfidensintervall og prediksjonsintervall

Litt repetisjon:

Har to variabler:	y	Responsvariabel	Eks	Lungekraft
	x	Forklaringsvariabel		Røyking
		Prediktor		Fyrstikker i lomma
		Kovariat		

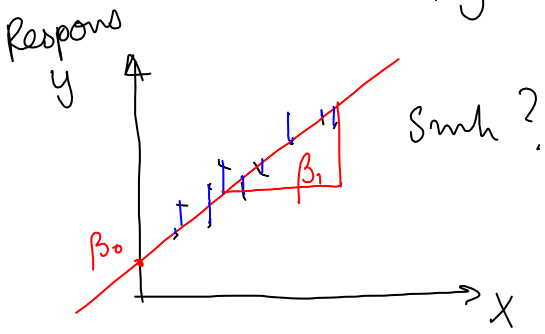


Smh?

Korrelasjonsanalyse: $H_0: \rho = 0$

Tester ved å bruke r.

Bivariat / symmetrisk



Smh?

Regresjonsanalyse: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

stigningstallet for linja:
Viser hvor mange enheter y forventes å øke, når x øker én enhet.

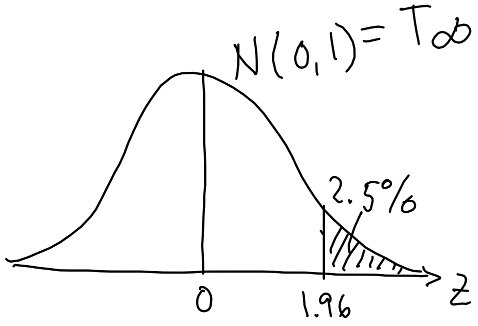
Tester for om det er en sammenheng: $H_0: \beta_1 = 0$

ved å bruke $\hat{\beta}_1$

95% Konfidensintervall for β_1 :

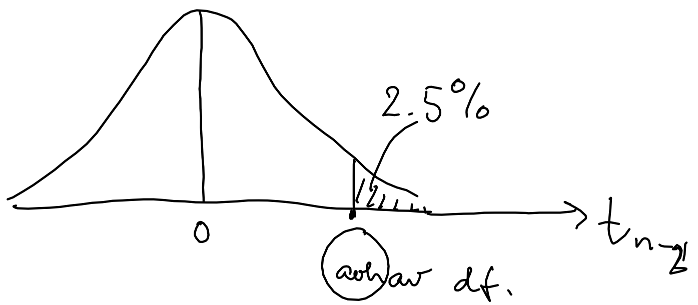
$\hat{\beta}_1 =$ (formel)

$$\frac{\hat{\beta}_1 - \beta_1}{S.E.(\hat{\beta}_1)} \sim T_{n-2}$$



$$\frac{\bar{X} - \mu}{\frac{sd}{\sqrt{n}}} \sim \cancel{N}\ T_{n-1}$$

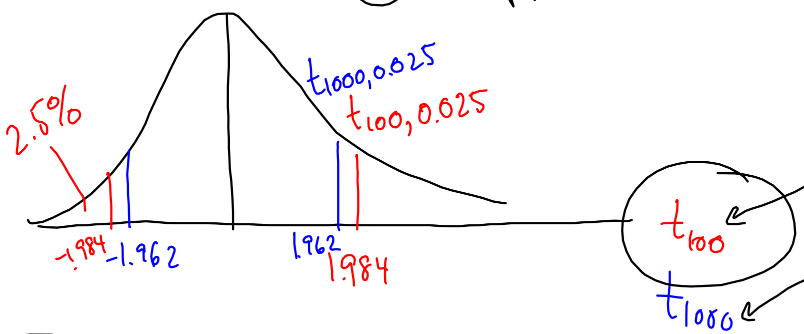
$SE(\bar{x}) \leftarrow \frac{sd}{\sqrt{n}}$



Her : $n = 153$

$df = n-2 = 151$

Table D



$$P\left(-1.984 < \frac{\hat{\beta}_1 - \beta_1}{S.E(\hat{\beta}_1)} < 1.984\right) = 0.95$$

$$P\left(-1.984 < \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} < 1.984\right) = 0.95$$

$$P\left(-2 < \frac{\hat{\beta}_1 - \beta_1}{SE} < 2\right) = 0.95$$

$$P\left(-2.626 < \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} < 2.626\right) = 0.99$$

$$P\left(\hat{\beta}_1 - 1.984 \cdot SE(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + 1.984 \cdot SE(\hat{\beta}_1)\right) = 0.95$$

-2.626 $+2.626$ 0.99

→ 95% KI for β_1 er (les fra R-utskrift): $1.7892 \pm 1.984 \cdot 0.2385$

$[1.32, 2.26]$ Intervallet inneholder ikke 0

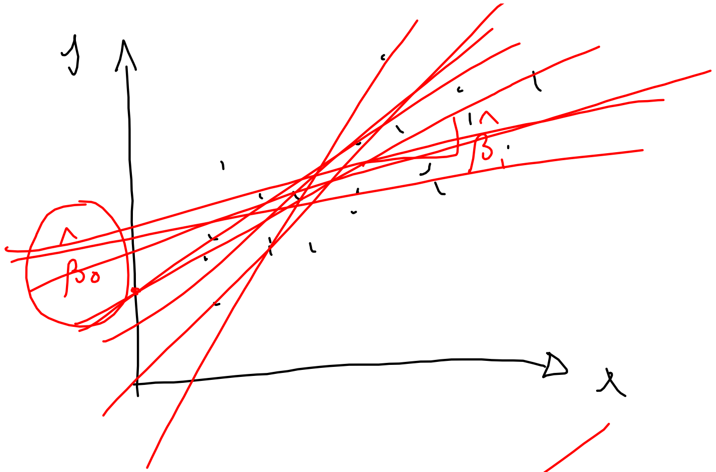
→ Forkast H_0 på nivå 0.05

→ 99% KI for β_1 er (les fra R)

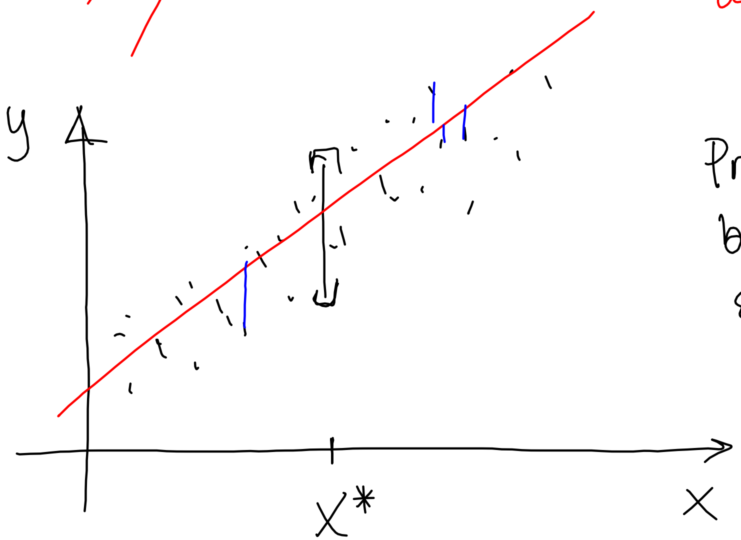
$1.7892 \pm 2.626 \cdot 0.2385$

$[1.15, 2.43]$ → inneholder ikke 0

→ Forkast H_0 på nivå 0.01



Konfidensintervallene for β_0 og β_1 viser oss usikkerheten i hvor regresjonslinja går.
Mer spesifikt: KI for β_1 viser oss usikkerheten i effektmålet, altså tallet for sammenhengen mellom x og y



Prediksjonsintervall er et intervall som beregnes for en spesifikk x^* verdi og som forteller oss hvor stor variasjon i yverdier vi kan forvente for denne x^*-verdien

Ligningen for sammenhengen (modellen) er:

$$y_i = \boxed{\beta_0 + \beta_1 \cdot x_i} + \varepsilon_i \qquad \varepsilon_i \sim N(0,1)$$

Modell: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0,1)$

$\underbrace{\text{Hva forventer vi at } y \text{ skal være,}}$ når $x = x^*$?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Hent $\hat{\beta}_0$ og $\hat{\beta}_1$ fra R-utskrift

Prediksjonsintervallet forteller om usikkerheten i $\hat{y}$, og er gitt ved:

$$\hat{y} \pm t_{0.025} \cdot SE(\hat{y})$$

$$SE(\hat{y}) = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$sd(x) = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

$$(sd(x))^2 \cdot (n-1) = \sum (x_i - \bar{x})^2$$

$$s = \frac{\sum (y - \hat{y})^2}{n-2}$$

Fra R-utskrift for summary(lm)
Residual standard error

$$SE(\hat{y}) = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(sd(x))^2 \cdot (n-1)}}$$

Fra deskriptiv statistikk for X