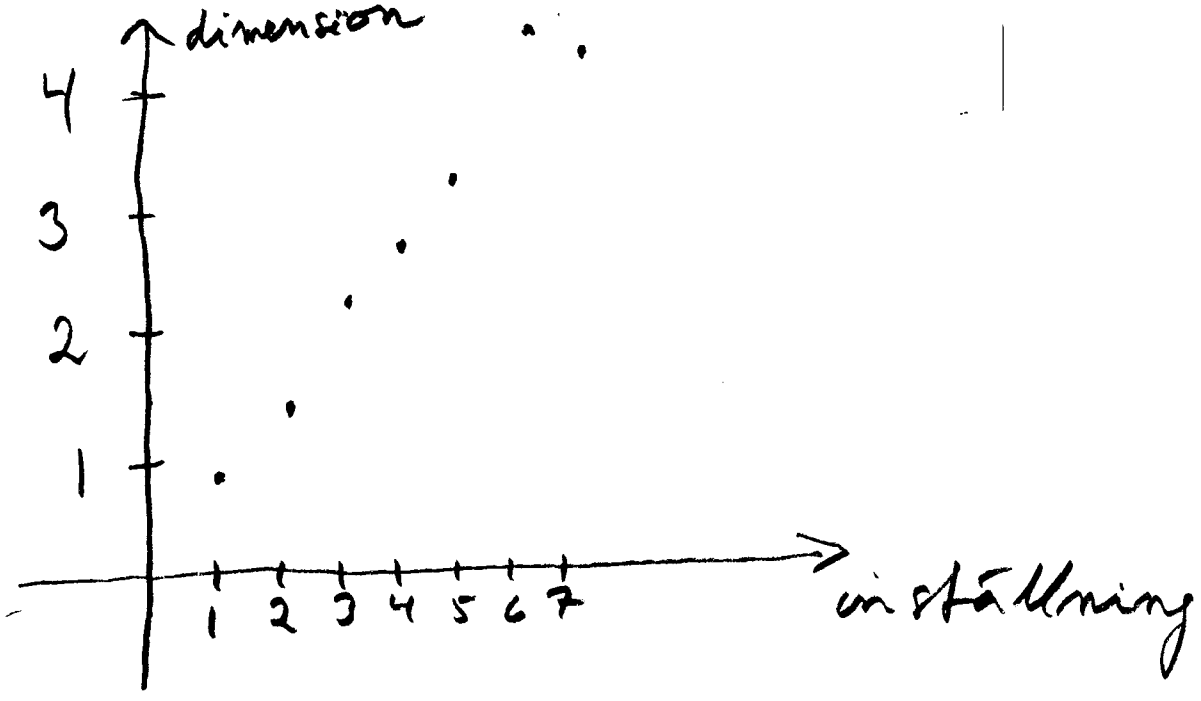
Regressionsanalys

Ex. En maskin tillverkar en viss vara vars dimension kan påverkas genom en inställningsratt på maskinen. Dimensionen beror dock inte exakt av inställningen utan där finns en viss slumpkomponent. Detta betyder att om man t.ex. tillverkar två exemplar av varan (med samma inställning) så får de inte nödvändigtvis samma dimension. Man tillverkade nu sju st varor med inställningarna 1, 2, 3, 4, 5, 6 och 7 respektive. Därvid blev de respektive dimensionerna 0.9, 1.4, 2.2, 2.7, 3.2, 4.3 och 4.2:



Som synes verkar dimensionen bero linjärt av inställningen (d.v.s. som på en viss slumpvariation) och detta kan man "fånga upp" genom modellen:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2 \cdots 7)$$

↑ n:r i F.S.

Här betecknar y_i dimensionen och x_i inställningen medan ε_i är slumpfaktorn. Man brukar anta att $\varepsilon_1, \varepsilon_2 \cdots \varepsilon_7$ är oberoende $N(0, \sigma)$.

Från de data vi har kan vi nu uppskatta β_0 (= det ställe där linjen skär y-axeln) och β_1 (= linjens lutning). I FS (s. 7) kallas dessa skattningar b_0 resp b_1. För att bestämma b_1 behöver vi $\sum x_i y_i = 1 \cdot 0.9 + \cdots + 7 \cdot 4.2 = 92.3$, $\sum x_i = 1 + \cdots + 7 = 28$, $\sum y_i = 0.9 + \cdots + 4.2 = 18.9$ och $\sum x_i^2 = 1^2 + \cdots + 7^2 = 140$. Vi får först

$$b_1 = \frac{92.3 - \frac{(28)(18.9)}{7}}{140 - \frac{(28)^2}{7}} \approx 0.5964,$$

(sedan $b_0 = \bar{y} - b_1 \cdot \bar{x} =$

$= \frac{18.9}{7} - 0.596 \cdot \frac{28}{7} = 0.3143$. Detta ger den s.k. regressionslinjen:

$$y = \underbrace{0.3143}_{\approx \beta_0} + \underbrace{0.5964}_{\approx \beta_1} x$$

Prova att rita in linjen i diagrammet. Som den ser "passar" den bra till punkterna. Dess exakta meningen av "passar" ovan är att linjen ligger så att summan av kvadraterna på (de lodräta) avstånden från punkterna till linjen är så liten som möjligt. Ju mindre denna minsta kvadratsumma är desto bättre passar linjen till punkterna.

Om vi inför $\hat{y}_i = b_0 + b_1 x_i$ (jfr. F.S) så kan denna minsta kvadrat-summa skrivas $\sum_{i=1}^{7} (y_i - \hat{y}_i)^2$ och man kan med hjälp av den skatta σ. Man har nämligen att $\sigma^2 \approx$

$\approx \frac{1}{7-2} \sum_{i=1}^{7} (y_i - \hat{y}_i)^2 = \frac{1}{7-2} [(0.9 - 0.3143 - 0.5964 \cdot 1)^2 +$

$+ \ldots + (4.2 - 0.3143 - 0.5964 \cdot 7)^2] = \frac{1}{5} \cdot 0.28 =$

$= 0.056$ varur $\sigma \approx 0.237$. Vi ska nu även beräkna ett konfidensintervall för β_1 på det vanliga sättet d.v.s

I_{β_1}: skattn $\pm$ t · medelfelet (jfr. F.S)

Att det blir t-fördelning här beror på att vi (i princip) har situationen med N-fördelning och okänt σ. Frihetsgraderna för t-fördelningen blir dock n-2 i stället för n-1. Medelfelet står inte i F.S. men blir

$$\frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \approx \frac{0.237}{\sqrt{(1-4)^2+\cdots+(7-4)^2}} =$$

$= 0.0448$. Med 95 % konfidensgrad får vi I_{β_1}: $0.5964 \pm t_{0.025}(5) \cdot 0.0448$ (95%)

$t_{0.025}(5) = 2.571$ enl tabell 8

$\Rightarrow 0.481 < \beta_1 < 0.711$ (95 %)

Eftersom 0 ej finns med i I_{β_1} så kan vi förkasta $H_0: \beta_1 = 0$ till förmån för $H_1: \beta_1 \neq 0$. (Nivå: 5 %)

Förklaringsgrad (= Koeff. of determination)

Ett mått på y_i-värdenas (totala) variation är $\sum_{i=1}^{n}(y_i-\bar{y})^2$. Det kan visas att denna totala variation kan delas upp i två summor på

Följande vis:

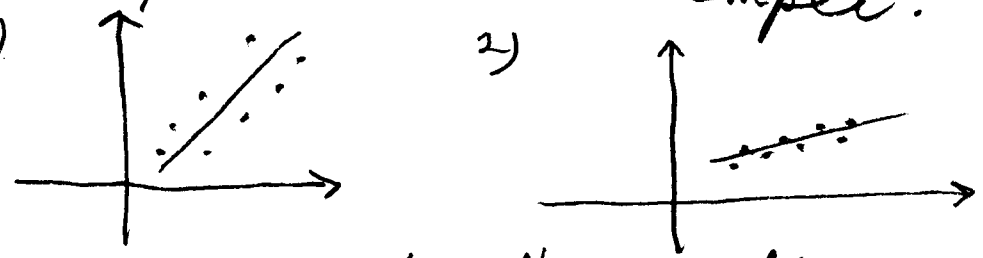
$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SST}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SSR}+\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SSE}$$

SSR är den summa som hör till regressionen medan SSE är den summa som hör till felet. För att regressionen ska vara bra bör SST ha ett värde nära SSR d.v.s. SSE bör vara liten. Förklaringsgraden (R^2) definieras nu som $\frac{SSR}{SST}=\left(1-\frac{SSE}{SST}\right)$ jfr. F.S.

Om man uppfattar både X och y-värdena som observationer på stokastiska variabler gäller $R=\pm$ korrelationskoeff(s) mellan X och Y. (Man väljer samma tecken som b_1 har.)

SSE = 0.28 (enl. tidigare räkning) SST = $=(0.4-2.7)^2+\dots+(4.2-2.7)^2=10.24$ och SSR = SST − SSE = 9.96 $\Rightarrow R^2=\frac{9.96}{10.24}\approx 97\%$ d.v.s. en mycket bra (modell, regressions) ($s=\pm(-)\,0.986$)

Observera att det inte är storleken på b_1 som avgör hur bra regressionsmodellen är utan hur bra anpassningen till linjen är. Ett exempel:



R^2 är mindre för 1) än för 2).

Multipel regression.

Om maskinen ovan har två mätare vars inställningar båda påverkar dimensionen så får man modellen: ($X1$ och $X2$ kallas för "prediktorer"*)

$$y_i = \beta_0 + \beta_1 (X1)_i + \beta_2 (X2)_i + \varepsilon_i \quad (i = 1, 2 \cdots n)$$

där $\varepsilon_1, \varepsilon_2 \cdots \varepsilon_n$ är oberoende $N(0, \tau)$.

Här måste skattningarna av β_0, β_1 och β_2 göras med hjälp av datorn som samtidigt testar (t.ex. genom p-värde) om ett eller flera β_j ($j > 0$; β_0 brukar sällan testas) är noll, d.v.s. om motsvarande prediktor inte behöver i modellen. Man kan också testa den allmänna hypotesen $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ med det s.k. F-test. En s.k. ANOVA tabell framräknad av datorn kan se ut såhär:

*) eller för "beroende variabler"

	DF	SS	MS
Regression:	4	84415.08 (SSR)	21103.77
+ Residual:	85	128276.4 (SSE)	1509.135
Total:	89	212691.5 (SST)	

(jfr. tabell på sid 502). Obs att
4 = antal prediktorer och $89 = n-1$
Vidare är $MS = SS/DF$ och 1509.135
en skattning av σ^2 i modellen:
"respons" eller "beroende variabel"

$$y_i = \beta_0 + \beta_1 (X1)_i + \beta_2 (X2)_i + \beta_3 (X3)_i + \beta_4 (X4)_i + \varepsilon_i$$

där $\varepsilon_1, \varepsilon_2 \dots \varepsilon_{90}$ är ober $N(0, \sigma)$

För att (på nivå 1% (t.ex.)) testa $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
mot H_1: minst ett $\beta_j \neq 0$ beräknar
vi $F = \frac{21103.77}{1509.135} = 13.98402$
och förkastar H_0 om $F >$
$> F_{0.01}(4, 85)$: Tabell 9b ger
$F(4, 85)$ = ett värde mellan 3.649
och 3.513 och vi ser att
H_0 förkastas med besked.

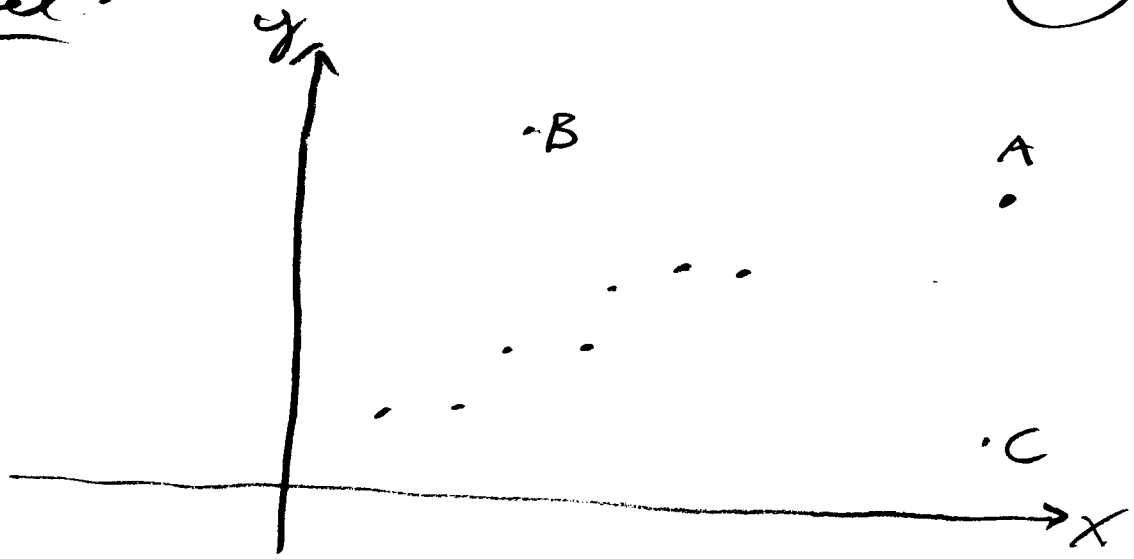
Några termer

1) Med "homoskedacitet" menas att
σ ovan är samma för alla ε_i.

2) Med "multikollinearitet" menas att det finns ett linjärt beroende mellan två eller flera av prediktorerna. Detta leder (ofta) till väldigt stora (varierande) medelfel (=SE = standard error i datautskriften). För de inblandade prediktorerna (x_j:na) gäller motsvarande β_j därför (i princip) ej att skatta.

3) Med "outlier" menas ett värde som ligger "utanför" de andra värdena. Det finns två slag av outliers, dels de som ligger utanför värdena i "prediktorrummet", dels de som ligger utanför värdena i "responsrummet".

Exempel:



A ligger utanför i prediktor-rummet (=har ett x-värde långt från de andra). B ligger utanför i responsrummet (=har ett y-värde långt från de andra). C ligger utanför "på båda sätten".
Det är ofta en svår fråga (att avgöra) om en outlier är en felaktig mätning eller ej.

Dummyvariabel

Låt oss gå tillbaka till det inledande exemplet, men anta att det finns två maskiner. Med hjälp av en variabel som kan ta två

värden (egentligen 0 och 1) kan man skriva upp en modell som innefattar båda maskinerna:

$$y_i = \beta_0 + \beta_1 (X1)_i + \beta_2 (X2)_i + \varepsilon_i \quad (i = 1, 2 \ldots n)$$

Här är (X1) i tillverkningen medan (X2) är = 1 om maskin 1 användes och = 0 om " 2 användes. Genom att t.ex. testa $H_0: \beta_2 = 0$ mot $H_1: \beta_2 \neq 0$ kan man undersöka om det spelar roll för dimensionen vilken maskin som användes.

χ^2-test

Ex. Pö 7 tal 1a.

$H_0: p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$, $p_4 = \frac{1}{16}$

ska testas mot H_1: Minst ett p_i skilt från värdet i H_0.

Detta är ett annorlunda test i den meningen att vi inte (som tidigare) testar en parameter utan en sannolikhetsfördelning,

Testvariabel (jfr. FS. sid 8)

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} \quad \text{där } E_i = n \cdot p_i$$

Här är O_1, O_2, O_3, O_4 de observerade värdena och E_1, E_2, E_3, E_4 de förväntade värdena om H_0 är sann. Om H_0 inte är sann bör O_i och E_i avvika från varandra och χ^2 därför bli förskjuten mot större värden. Divisionen med E_i har skett för att testvariabeln ska få en känd fördelning, i detta fall en χ^2-fördelning med $n-1$ frihetsgrader. (Nivå: 5%) Utfallet av test:

$$\frac{\left(57 - \frac{9}{16}\cdot 100\right)^2}{\frac{9}{16}\cdot 100} + \ldots + \frac{\left(11 - \frac{1}{16}\cdot 100\right)^2}{\frac{1}{16}\cdot 100} =$$

$$= 6.13 < \chi^2_{0.05}(3) = 7.815 \quad \left(\text{enl. tabell } \tfrac{7a}{}\right)$$

$\Rightarrow H_0$ förkastas ej.

I detta fall hade vi en helt känd sannolikhetsfördelning. Har man inte det minskas frihetsgraderna med ett för varje parameter i fördelningen man tvingas skatta. Detta är lite av överkurs, den som vill kan gå igenom exempel 14.4. på sid 610. (Obs. att skattningen av λ är $\frac{1}{202}(156 \cdot 0 + 63 \cdot 1 + 29 \cdot 2 + 14 \cdot 3)$ där man alltså räknat > 2 som 3.)

Test av oberoende

Vi går igenom exemplet på sid 616. Nu används (se F.S. sid 7)

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{med } E_{ij} = \frac{R_i C_j}{n}$$

som testvariabel. Här är R_i = i:te radens summa och C_j = j:te kolonnens summa. Om bilmärke och det sätt man ser på bilmärket ("sportigt" eller säkert) är oberoende av varandra

så är testvariabeln X^2-fördelad med $(r-1)(c-1) = (3-1)(2-1) = 2$ frihetsgrader, men finns där ett beroende så blir (den) förkastelsen mot större värden. Tabellen förklarad:

Bilmärke	Sportigt	Säkert	TOTAL
BMW	256	74	330 ($=256+74$) ← R_1
Mercedes	41	42	83 ← R_2
+ Lexus	66	34	100 ← R_3
Total	363	150	513 ← n
	↑ C_1	↑ C_2	

T.ex. $E_{12} = \frac{330 \cdot 150}{513} = 96.5$

att jämföra med $O_{12} = 74$

Vi får $X^2 = \frac{(256 - 233.5)^2}{233.5} +$

$+ \frac{(74 - 96.5)^2}{96.5} + \ldots + \frac{(34 - 29.2)^2}{29.2} =$

$= 26.8 > X^2_{0.05}(2) = 5.991$ (enl. tabell 7a) H_0 förkastas alltså på nivå 5 %. (Ja, t.o.m. på nivå 0.1 % förkastas H_0 ty $X^2_{0.001}(2) = 13.816$)