

Föreläsning 13: Multipel Regression

Matematisk statistik

David Bolin
Chalmers University of Technology
Oktober 15, 2018



Ett exempel från fysikalisk kemi

- Hastigheten för en kemisk reaktion ökar ofta då temperaturen höjs. En grov tumregel är att hastigheten fördubblas för varje ökning av temperaturen med 10 K.
- Hastighetskonstantens temperaturberoende kan för en reaktion i de flesta fall modelleras med Arrhenius' ekvation

$$k = Ae^{-E_a/(RT)}$$

där E_a är den så kallade (skenbara) aktiveringsenergin, R är den allmänna gaskonstanten och T är temperaturen.

- Aktiveringsenergin kan bestämmas experimentellt genom att man mäter k vid ett antal temperaturer och ansätter:

$$\log(k_i) = \log(A) - \frac{E_a}{R} \frac{1}{T_i} + \varepsilon_i$$

- Med $Y_i = \log(k_i)$ och $x_i = T_i^{-1}$ ser vi att detta är en regressionsmodell!

Enkel linjär regression

Vi har gjort mätningar av en responsvariabel Y för fixerade värden på en förklarande variabel x , som kan väljas utan fel.

Vi ansätter en linjär modell för $(Y_i, x_i), i = 1, \dots, n$:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (1)$$

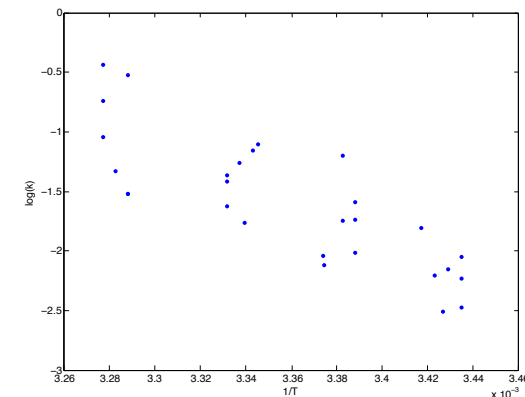
där ε_i är oberoende $N(0, \sigma^2)$ slumpvariabler som beskriver mätfelet och β_0 och β_1 är parametrar som beskriver det linjära sambandet.

Ett annat sätt att skriva upp modellen på är alltså att

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2),$$

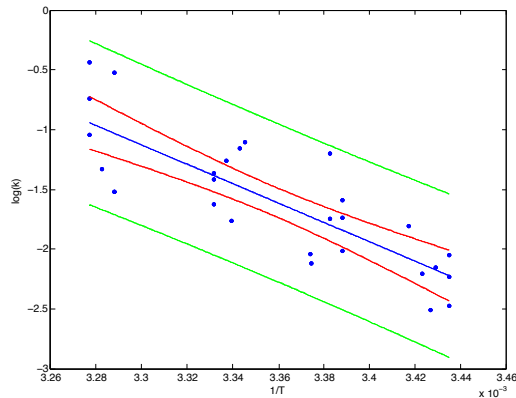
det vill säga att vi har ett linjärt samband som bestämmer väntevärdet hos Y och en viss mätfelsvarians σ^2 som beskriver de enskilda observationernas variation kring väntevärdet $\beta_0 + \beta_1 x$.

Mätningar från laborationer 2014



I kursen Fysikalisk kemi ingår en laboration om bestämning av hastighetskonstant för reaktionen mellan väteperoxid och jodidjon. Ovan ses datan från alla sju grupper från höstterminen 2014.

Resultat



Baserat på datan får vi $\beta_0^* = \log(A)^* = 25.6260$ samt $\beta_1^* = \frac{E_a^*}{R} = -8107$. Alltså är $A^* = 1.34 \cdot 10^{11}$ och $E_a^* = 67 \text{ kJ/mol}$.

Multipel linjär regression

Vi har gjort mätningar av en responsvariabel Y för fixerade värden på förklarande variabler $x_1, \dots, x_p$, som kan väljas utan fel.

Vi ansätter en linjär modell för $(Y_i, x_{1,i}, \dots, x_{p,i}), i = 1, \dots, n$:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \varepsilon_i \quad (2)$$

där ε_i är oberoende $N(0, \sigma^2)$ slumpvariabler som beskriver mätfelet och β_i är parametrar som beskriver beroendet.

Vi kan formulera modellen på matrisform som

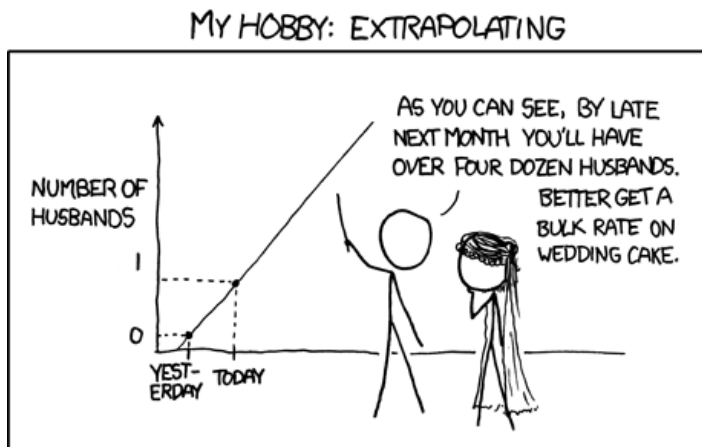
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

där

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{pmatrix} \quad \mathbf{E} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

Extrapolering

En mycket viktig komponent i en regressionsanalys är validering av modellen, det är också viktigt att inse att vi inte kan vara säkra på att den linjära modellen stämmer för x utanför mätområdet.



Parameterskattning

Sats

Givet att $\mathbf{X}^T \mathbf{X}$ är inverterbar fås minstakvadrat-skattningen av $\boldsymbol{\beta}$ som $\boldsymbol{\beta}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$. Skattningarna är väntevärdesriktiga och vi har

$$V(\beta_i^*) = \sigma^2 (\text{diagonalelement nummer } i+1 \text{ i } (\mathbf{X}^T \mathbf{X})^{-1}).$$

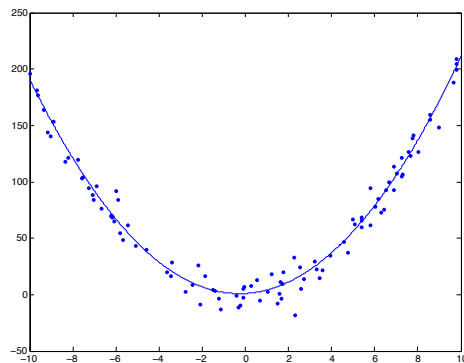
Vidare är $s^2 = Q_0 / (n - p - 1)$ där

$$Q_0 = \sum_{i=1}^n (y_i - \beta_0^* - \beta_1^* x_{1,i} - \dots - \beta_p^* x_{p,i})^2 = \mathbf{Y}^T \mathbf{Y} - \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y}.$$

Sats

Med $\mu_Y^*(\mathbf{x}_0) = \beta_0^* + \beta_1^* x_{0,1} + \dots + \beta_p^* x_{0,p}$ är $V(\mu_Y^*(\mathbf{x}_0)) = \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0$

Polynomregression

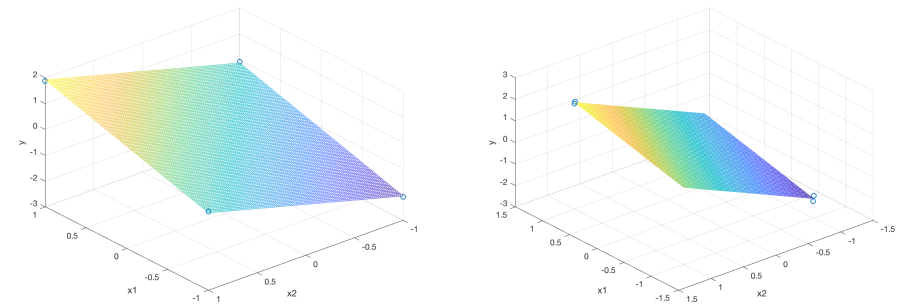


- Vi har gjort mätningar av en responsvariabel Y för fixerade värden på en förklarande variabel x . Ansätt modellen

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_p x_i^p + \varepsilon_i \quad (3)$$

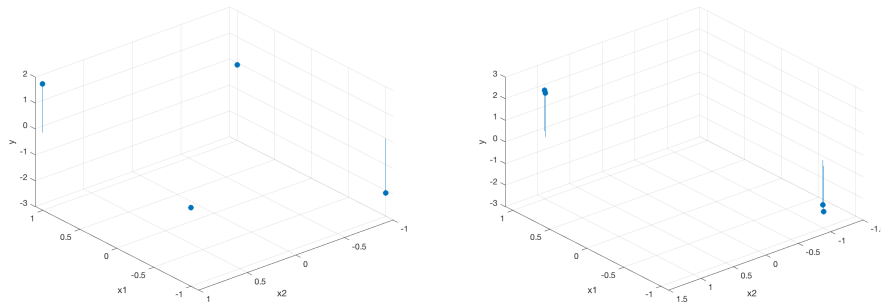
- Specialfall av multipel regresssion med $x_k = x^k$.

Kolinearitet



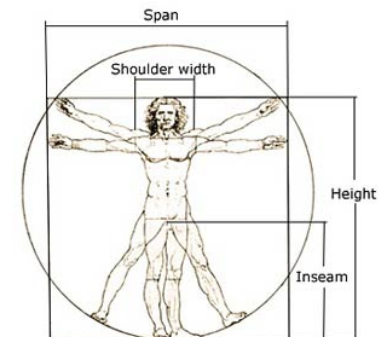
- Om alla observationer nästan ligger på en linje i (x_1, x_2) -planet blir skattningen av planets lutning i riktningen ortogonalt mot linjen mycket osäker.

Kolinearitet

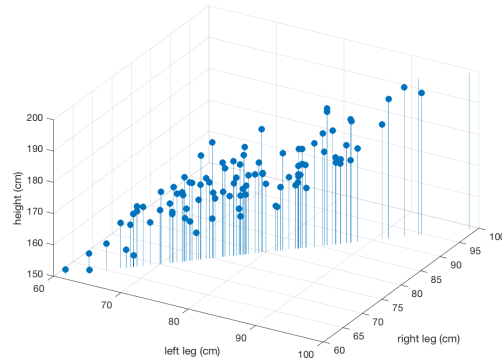


- Även om två förklarande variabler inte är linjärt beroende så kan vi få att kolumnerna i $\mathbf{X}$ blir linjärt beroende på grund av hur vi gör mätningarna.
- Om alla observationer ligger på en linje i (x_1, x_2) -planet blir kolumnerna i $\mathbf{X}$ linjärt beroende.

Exempel: skattning av längd



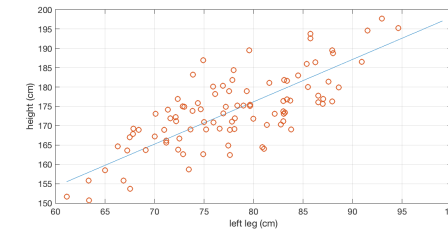
- Om vi vet hur långa ben en person har bör vi kunna göra en bra skattning av personens längd.
- Gör mätningar av personers längd samt benlängder för att kunna skatta ett regressions samband.



- Låt Y vara personens längd, x_ℓ längden av det vänstra benet, och x_r längden av det högra. Ansätt modellen

$$Y_i = \beta_0 + \beta_1 x_{\ell,i} + \beta_2 x_{r,i} + \varepsilon_i$$

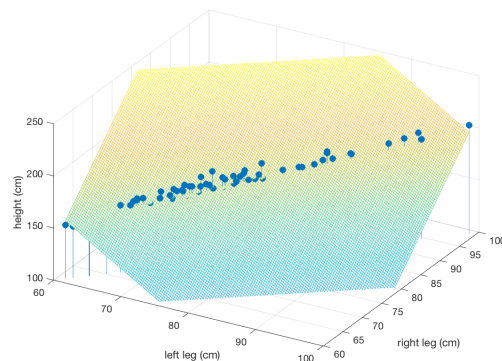
Vad gick fel?



- Det finns ett tydligt samband, men på grund av kolinearitet mellan variablerna kan modellen inte avgöra om det är vänster ben, höger ben, eller en kombination som förklarar längden.
- Lösning: Använd bara vänster (eller höger) ben som prediktor:

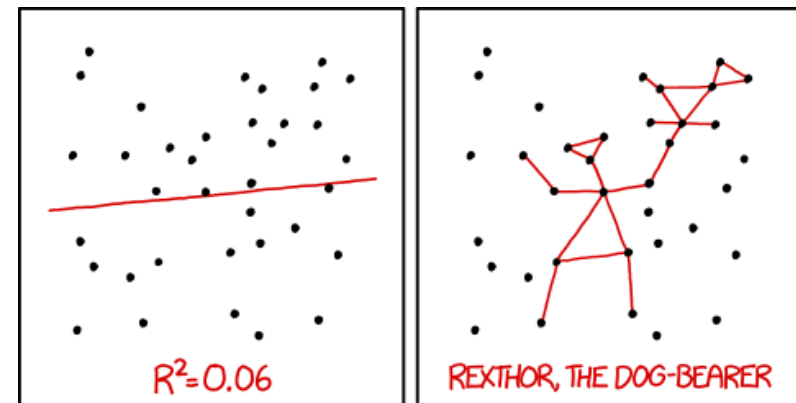
$$Y_i = \beta_0 + \beta_1 x_{\ell,i} + \varepsilon_i$$

- $\beta_1^* = 1.0953$, $I_{\beta_1} = (0.9283, 1.2624)$.



- Skattningar: $\beta_0^* = 88.6171$, $\beta_1^* = -3.9840$, $\beta_2^* = 5.0790$.
- Konfidensintervall
 $I_{\beta_1} = (-52.58, 44.62)$, $I_{\beta_2} = (-43.52, 53.67)$.
- Märkliga skattningar och ingen signifikant parameter!

Förklaringsgrad



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Korrelation och kausalitet

- Korrelation och regression säger inget om orsakssamband.
- Exempel: Dagar med hög glassförsäljning tenderar att ha fler drunkningsolyckor. Dags att förbjuda glass?
- Kausala frågor handlar om mekanismer som genererar datan eller prediktioner om vad som kommer ske efter någon åtgärd.
- Tex: Förändras antalet drunkningsolyckor om vi förbjuder glass?
- Det finns mängder av kausala påståenden i nyheterna!
 - "Ät frukost om du vill minska din risk för kranskärslssjukdomar"
- Vi måste vara försiktiga med kausala påståenden:
 - Om personer som äter frukost tenderar ha färre kranskärslssjukdomar, borde vi därför äta frukost?

Ett klassiskt exempel

Betrakta följande hypotetiska data angående effekterna av ett utbildningsprogram på åter-arresteringar hos före detta fångar:

	Arresterade	Ej arresterade	Frekvens
Medverkande	100	400	20%
Ej medverkande	500	500	50%

Typiska frågor vi kan vilja få svar på:

- Kan vi skatta risken för att en fånge återfaller i brott givet att den medverkade i programmet? (Icke-kausal)
- Minskar programmet frekvensen för återfall i brott? (Kausal)
- Vad vore återfallsfrekvensen om programmet vore obligatoriskt för fångar? (Kausal)

Vilka frågor kan vi svara på?

- Det beror på hur datan samlats in!

Randomiserade studier

- I en randomiserad studie slumpas det på förhand vilken grupp varje försöksperson ska tillhöra.
- Om datan för fångarna samlades in via en randomiserad studie, kan vi då svara på om programmet minskar återfallsfrekvensen?
- Antagligen ja! Men vi kan inte vara helt säkra, tex kan fångvaktarna bete sig olika mot de två olika grupperna.
- Om möjligt bör experimentet göras "double blind"
 - Försökspersonerna vet ej vilken grupp de tillhör (motverka placeboeffekt)
 - Forskarna (och fångvaktarna) vet ej vilken grupp varje försöksperson tillhör (motverka bias hos forskarna)
- Det är inte alltid möjligt, eller etiskt försvarbart, att randomisera
 - Vi kan till exempel inte randomisera vilka som ska röka om vi undersöker effekten rökning har på cancer!

Observationsstudier och confoundingvariabler

- I en observationsstudie sker urvalet hos de två grupperna genom att försökspersonerna själva väljer vilken grupp de ska tillhöra. Forskarna observerar endast resultaten.
- Om datan för fångarna samlades in via en observationsstudie, kan vi svara på om programmet minskar återfallsfrekvensen?
 - Nej! skillnaderna kan bero på systematiska skillnader i beteendet hos de som väljer att delta.
 - Tex vilka typer av brott de begått eller deras motivation att återintegreras i samhället.
- Motivation hos fångarna är ett exempel på en confoundingvariabel, dvs en faktor som
 - Influera utfallet hos försöket och är relaterad till den förklarande variabeln.
- Om båda dessa villkor är uppfyllda kan vi inte skilja effekten från den förklarande variabeln från confoundingvariabeln.
 - Vi kan inte förklara vad som orsakar effekten i utfallet!

Confoundingvariabel för fängelseexemplet

Motivation till återanpassning till samhället är en confoundingvariabel:

- Den influerar chansen att återfalla i brott.
- Den är relaterad till deltagande i utbildningsprogrammet, eftersom folk som deltar är motiverade.

Vi har alltså att

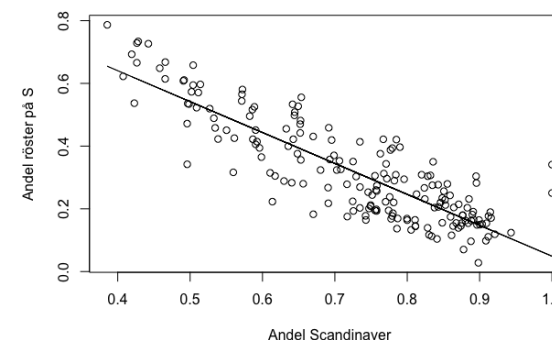
- De fångar som är motiverade och deltog i programmet återfaller sällan i brottslighet.
- De fångar som inte är motiverade deltog inte i programmet och återfaller ofta i brottslighet.

Vi kan alltså inte veta om återfallsfrekvensen beror på motivation eller deltagande, dessa två effekter mixas samman (confounded på engelska)

Analys av data från observationsstudier

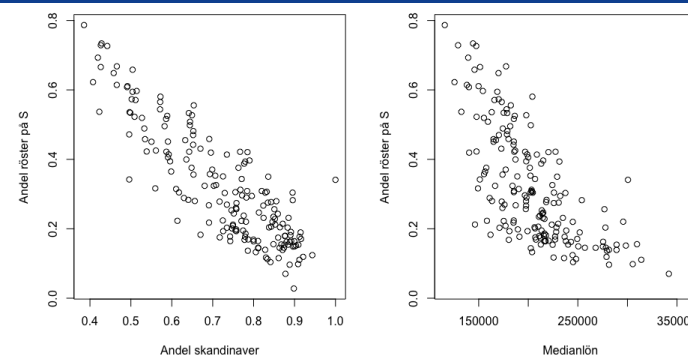
- Finns det något vi kan göra för att analysera kausala frågor med hjälp av data från observationsstudier?
- Jo, men det kräver mer arbete!
- Blockindelning/stratifiering: Ett sätt är att undersöka homogena delpopulationer och göra jämförelser inom varje delpopulation.
 - Jämför motiverade och omotiverade fångar för sig.
- Ett problem är att det kan vara svårt att veta att vi kontrollerat för varje relevant confoundingvariabel.
- Ett annat problem är att det kan vara svårt att få tillgång till data för confoundingvariabler.

Socialdemokrater och Skandinaver



- Datan visar alla valdistrikt i Malmö.
- X = andel av populationen som kommer från Skandinavien
- Y = andel av populationen som röstade på Socialdemokraterna i förra valet.
- $\rho = -0.8591938$, $R^2 = 0.73$.

Medianlön en confounder?



- Ta med även medianlön i distrikten som förklarande variabel i en multipel regressionsmodell.

$$Y = \beta_0 + \beta_1 X_{scand} + \beta_2 X_{salary} + \varepsilon$$

- $\beta_1^* = -0.89$, fortfarande signifikant ($p < 2 \cdot 10^{-16}$).
- $R^2 = 0.75$.

Modellval

- Som vi sett är en viktig fråga är hur vi väljer vilka förklarande variabler som ska ingå i en multipel regressionsmodell.
- För att undersöka om x_p bör tas med kan vi utföra hypotestest

$$H_0 : \beta_p = 0$$

$$H_1 : \beta_p \neq 0$$

genom att bilda teststorheten $\beta_p^*/d(\beta_p^*)$ med kritiskt område $C_\alpha = \{T : |T| > t_{\alpha/2}(n - p - 1)\}$.

Stegvis regression (I)

Låt $(\beta_0, \beta_i, \beta_j)$ beteckna modellen $Y = \beta_0 + \beta_i x_i + \beta_j x_j + \varepsilon$ och låt $Q_0(\beta_0, \beta_i, \beta_j)$ vara residualvariationen för modellen.

Steg 1

Välj (β_0, β_{i_1}) där $i_1 = \arg \min_i Q_0(\beta_0, \beta_i)$. Om β_{i_1} är signifikant, gå till Steg 2a, annars välj (β_0) som modell och avsluta modellval.

Steg 2a

Välj $(\beta_0, \beta_{i_1}, \beta_{i_2})$ där $i_2 = \arg \min_i Q_0(\beta_0, \beta_{i_1}, \beta_i)$. Om β_{i_2} är signifikant, gå till Steg 2b, annars välj (β_0, β_{i_1}) som modell och avsluta modellval.

Steg 2b

Testa om β_{i_1} är signifikant i $(\beta_0, \beta_{i_1}, \beta_{i_2})$. Behåll modellen om så är fallet, annars välj (β_0, β_{i_2}) . Gå till steg 3a.

Stegvis regression (II)

Fortsätt iterera stegen i modellvalet:

Steg m:a

Aktuell modell är $(\beta_0, \beta_{i_1}, \dots, \beta_{i_k})$. Välj $(\beta_0, \beta_{i_1}, \dots, \beta_{i_k}, \beta_{i_{k+1}})$ där $i_{k+1} = \arg \min_i Q_0(\beta_0, \beta_{i_1}, \dots, \beta_{i_k}, \beta_i)$. Om $\beta_{i_{k+1}}$ är signifikant, gå till Steg m:a, annars välj $(\beta_0, \beta_{i_1}, \dots, \beta_{i_k})$ som modell och avsluta modellval.

Steg m:b

Testa om någon parameter β_j är icke-signifikant i $(\beta_0, \beta_{i_1}, \dots, \beta_{i_k}, \beta_{i_{k+1}})$. Om så är fallet, välj bort den med lägst värde på teststorheten, annars behåll alla parametrar. Gå till Steg $(m + 1) : a$.

Massignifikans

- Vi måste vara försiktiga när vi gör försök som kräver flera tester.
- Till exempel är stegvis regression lite problematiskt eftersom valet av α i testerna i viss mån godtyckligt eftersom vi gör upprepade tester.
- Detta brukar kallas *massignifikansproblemet*:

