Föreläsning 13: Multipel Regression

Matematisk statistik

David Bolin
Chalmers University of Technology
Oktober 12, 2015

Modellbeskrivning

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 $(Y_i, x_i), i = 1, \ldots, n$:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

där ε_i är oberoende $\mathsf{N}(0, \sigma^2)$ slumpvariabeler 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

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

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 $\beta_0 + \beta_1 x$.

Minsta-kvadratmetoden

Antag den lite mer generella modellen att väntevärdet för Y ges av en funktion f:

$$\mathsf{E}(Y_i) = f(\theta_1, \ldots, \theta_k, x_i)$$

Parametrarna $\theta_1, \ldots, \theta_k$ skattas enligt minsta-kvadratmetoden genom att minimera kvadratfelet för den datan vi har

$$S(\theta_1, \ldots, \theta_k) = \sum_{i=1}^{n} (y_i - f(\theta_1, \ldots, \theta_k, x_i))^2$$

med avseende på parametrarna $\theta_1, \ldots, \theta_k$.
Lösningen brukar fås som lösningen till ekvationerna

$$\frac{\partial S}{\partial \theta_i} = 0, \quad \text{för } i = 1, \ldots, k$$

Ekvationssystemet löses av $\theta_1^*, \ldots, \theta_k^*$ som kallas för MK-skattningarna av $\theta_1, \ldots, \theta_k$.

MK-skattningar av β_0 och β_1

Inför

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Vi kan nu skriva lösningen till ekvationssystemet som

$$\beta_1^* = S_{xy}/S_{xx}$$
$$\beta_0^* = \bar{y} - \beta_1^*\bar{x}$$

MK-skattning av σ^2

Variansparametern σ^2 beskriver spridningen kring linjen och skattas som $s^2 = \frac{Q_0}{n-2}$ där

$$Q_0 = \sum_{i=1}^{n} (y_i - \beta_0^* - \beta_1^* x_i)^2.$$

Om vi räknar för hand kan vi använda att

$$Q_0 = S_{yy} - \beta_1^* S_{xx} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Vi delar på $n-2$ eftersom två frihetsgrader används till att skatta β_0 och β_1.

Sats

Vi har att $\mathsf{E}(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$ och $\mathsf{V}(\bar{Y}) = \frac{\sigma^2}{n}$. Dessutom är

$$\mathsf{E}(\beta_1^*) = \beta_1 \qquad \mathsf{V}(\beta_1^*) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})}$$

Vi ser alltså att β_1^* är väntevärdesriktig.

Sats

Med $\mu_Y^*(x_0) = \beta_0^* + \beta_1^* x_0$ är

$$\mathsf{E}(\mu_Y^*(x_0)) = \beta_0 + \beta_1 x_0$$

samt

$$\mathsf{V}(\mu_Y^*(x_0)) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

Med $x_0 = 0$ ser vi β_0^* är väntevärdesriktig.

Skattningarnas fördelningar

Sats

Om ε_i är normalfördelade gäller att $\bar{Y}$, β_0^*, β_1^* och $\beta_0^* + \beta_1^* x_0$ också är normalfördelade.

Eftersom skattarna är summor av Y_i så gäller enligt CGS satsen approximativt även om ε_i avviker från normalfördelningen.

Sats

Om ε_i är normalfördelade så gäller att

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$$

vidare är s^2 oberoende av $\bar{Y}$, β_0^*, β_1^* och $\beta_0^* + \beta_1^* x_0$.

Konfidensintervall och test

Låt θ vara någon av β_0, β_1 eller $\beta_0 + \beta_1 x_0$. Vi vet att dess skattningar är normalfördelade och vi har tagit fram variansen av skattningarna. Låt $d(\theta^*)$ vara standardavvikelsen för skattningen, vi har då att

$$T = \frac{\theta^* - \theta}{d(\theta^*)} \sim t(n-2)$$

vilken på vanligt sätt används för att göra test och bilda konfidensintervall,

$$I_\theta = (\theta^* \pm t_{\alpha/2}(n-2)d(\theta^*))$$

Konfidensintervall

- Konfidensintervall för β_0:

$$I_{\beta_0} = \left(\beta_0^* \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \right)$$

- Konfidensintervall för β_1:

$$I_{\beta_1} = \left(\beta_1^* \pm t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}} \right)$$

- Konfidensintervall för $\mu_Y(x_0) = \beta_0 + \beta_1 x_0$:

$$I_{\mu_Y(x_0)} = \left(\beta_0^* + \beta_1^* x_0 \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

Prediktionsintervall

- Ibland vill man veta var en framtida observation kommer vara för ett visst värde på x, då används ett *prediktionsintervall* .
- Skillnaden mellan ett prediktionsintervall $I_{Y(x_0)}$ och ett konfidensintervall $I_{\mu_Y(x_0)}$ är att $I_{\mu_Y(x_0)}$ talar om var väntevärdet troligen ligger, medan $I_{Y(x_0)}$ talar om var en framtida observation troligen ligger.
- Eftersom observationerna har en viss spriding kring linjen så måste prediktionsintervallet vara bredare än konfidensintervallet, och man kan visa att

$$Y^*(x_0) \sim \mathsf{N}\left(\beta_0 + \beta_1 x_0, \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})\right).$$

- Ett prediktionsintervall bildas nu som

$$I_{Y(x_0)} = \left[\beta_0^* + \beta_1^* x_0 \pm t_{p/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right]$$

Modellvalidering

En mycket viktig komponent i en regressionsanalys är validering av modellen, vilket betyder att vi måste försäkra oss om att det är lämpligt att ansätta en enkel regressionsmodell. Det vanligaste sättet att göra detta på är att beräkna de så kallade residualerna

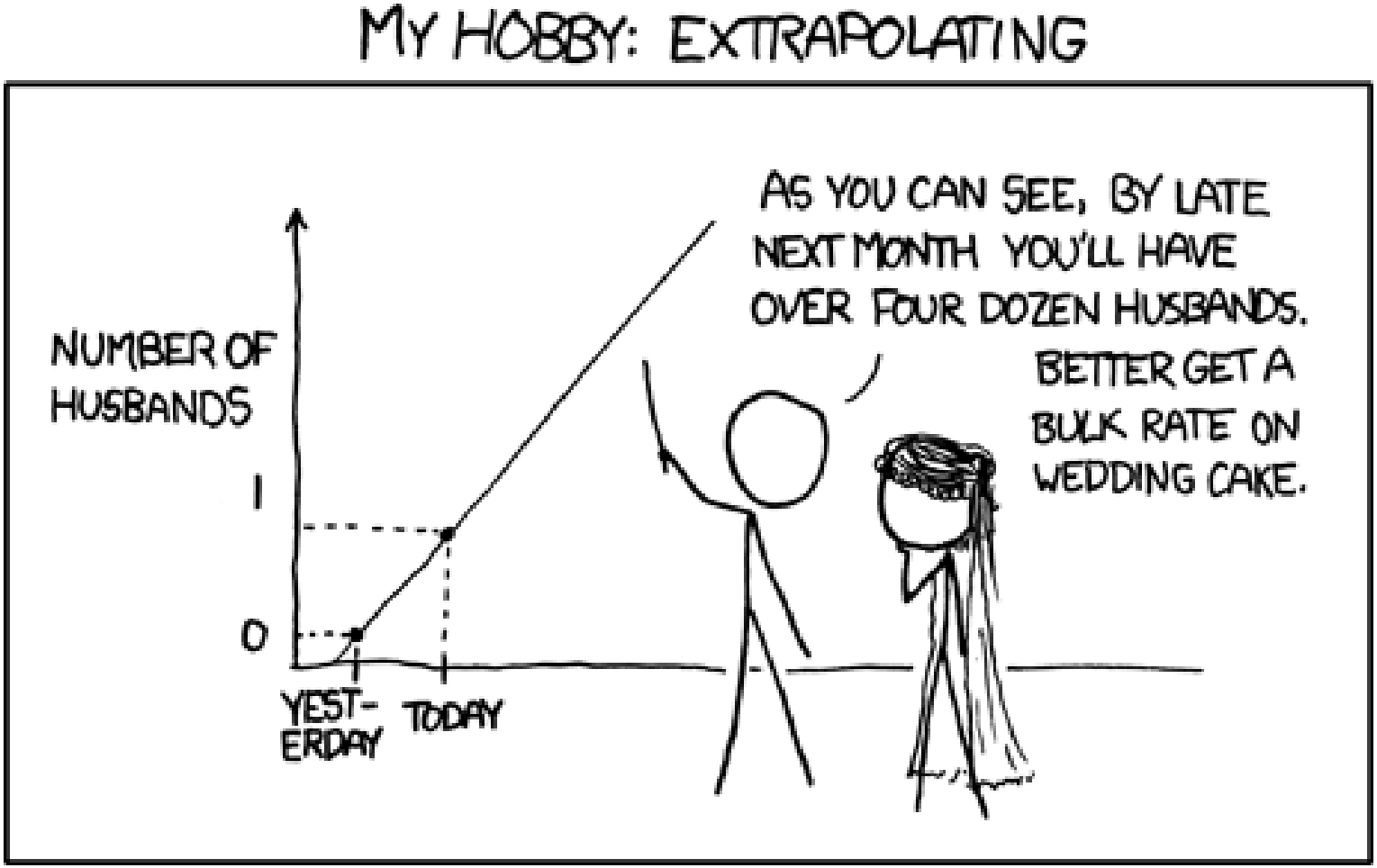
$$e_i = y_i - \beta_0^* - \beta_1^* x_i$$

Om modellen är korrekt bör dessa residualer

- vara ungefär normalfördelade med väntevärde noll.
- inte uppvisa någon speciell struktur som funktion av x.
- ha ungefär samma variation för alla olika värden på x, vi får till exempel inte ha att variansen verkar vara större för stora värden på x.

Undersök visuellt genom att plotta residualerna som funktion av x och använda normalfördelningsdiagram.

Det är också viktigt att inse att vi inte kan vara säkra på att den linjära modellen stämmer för x utanför mätområdet.



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

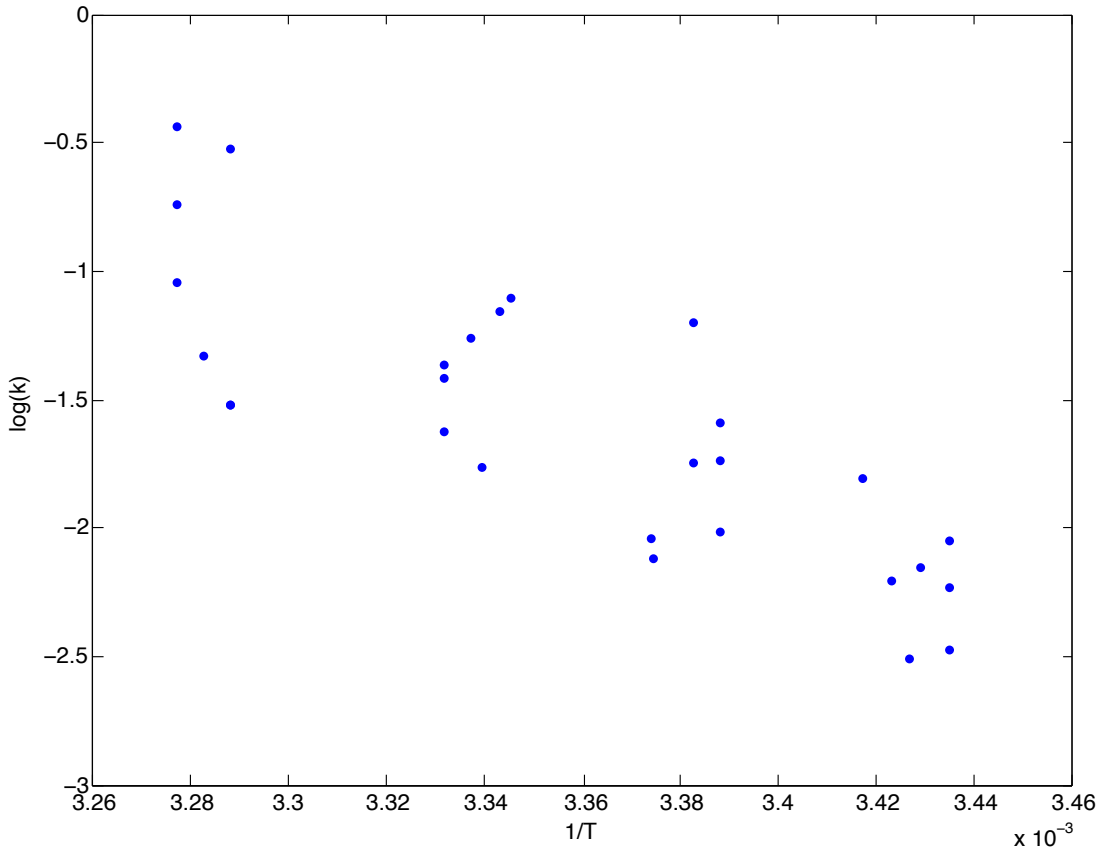
$$k = Ae^{-E_a/(RT)}$$

  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:

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

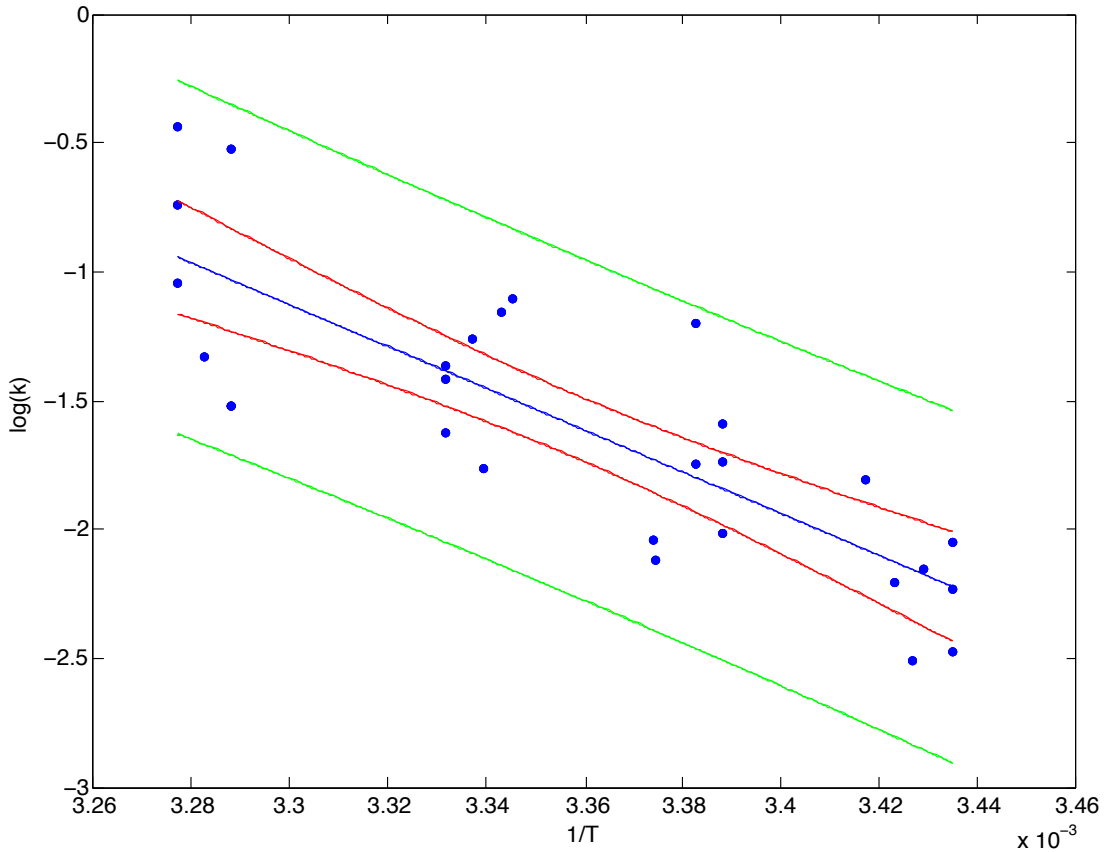
- Med $Y_i = \log(k_i)$ och $x_i = T_i^{-1}$ ser vi att detta är en regressionsmodell!

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 $E_a^* = 67 kJ/mol$.