Föreläsning 12: Regression

Matematisk statistik

David Bolin
Chalmers University of Technology
Maj 15, 2014

Binomialfördelningen

Låt $X \sim \text{Bin}(n, p)$. Vi observerar x och vill ha information om p.

$p^* = x/n$ är en väntevärdesriktig skattning av p med varians $\text{V}(p^*) = p(1-p)/n$.

Centrala gränsvärdessatsen som säger att X är approximativt $\text{N}(np, np(1-p))$-fördelad om n är stor. En tumregel brukar vara att om $np(1-p) > 10$ så fungerar normalapproximation bra. Vi har då att p^* är approximativt $\text{N}(p, p(1-p)/n)$-fördelad och att

$$\frac{p^* - p}{\sqrt{p^*(1-p^*)/n}}$$

är approximativt $\text{N}(0, 1)$-fördelad.

Konfidensintervall

Ett konfidensintervall med approximativ konfidensgrad $1-\alpha$ fås som

$$I_p = [p^* \pm z_{\alpha/2}\sqrt{p^*(1-p^*)/n}]$$

Eftersom p är en sannolikhet bildas ensidiga intervall som

$$[0, p^* + z_{\alpha}\sqrt{p^*(1-p^*)/n}]$$

och

$$[p^* - z_{\alpha}\sqrt{p^*(1-p^*)/n}, 1]$$

Hypotestest

Vill vi testa

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0$$

eller en ensidig hypotes $H_1 : p > p_0$ eller $H_1 : p < p_0$.

Under H_0 är $X \sim \text{Bin}(n, p_0)$, och vi kan direkt beräkna p-värdet för testet:

- $p = \text{P}_{H_0}(X \geq x)$ om $H_1 : p > p_0$.
- $p = \text{P}_{H_0}(X \leq x)$ om $H_1 : p < p_0$.
- $p = 2\text{P}_{H_0}(X \geq x)$ om $x \geq np_0$ och $p = 2\text{P}_{H_0}(X \leq x)$ om $x \leq np_0$ i fallet $H_1 : p \neq p_0$.

Jämförelser i binomialfördelningen

Antag att $X_1 \sim \text{Bin}(n_1, p_1)$ och $X_2 \sim \text{Bin}(n_2, p_2)$ och vi vill testa om det är rimligt att anta att $p_1 = p_2$.
$p_1^* - p_2^*$ är en väntevärdesriktig skattning av $p_1 - p_2$.
Om $n_1 p_1(1-p_1)$ och $n_2 p_2(1-p_2)$ båda är stora (säg större än 10) så är $p_1^* - p_2^*$ approximativt normalfördelad. Vi har

$$\text{V}(p_1^* - p_2^*) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

ett konfidensintervall för $p_1 - p_2$ fås som

$$I_{p_1-p_2} = \left(p_1^* - p_2^* \pm z_{\alpha/2} \sqrt{\frac{p_1^*(1-p_1^*)}{n_1} + \frac{p_2^*(1-p_2^*)}{n_2}} \right)$$

Hypotestest

Om vi vill testa $H_0 : p_1 = p_2$ kan vi använda en poolad variansskattning eftersom vi har att under H_0 så gäller

$$\mathsf{V}_{H_0}(p_1^* - p_2^*) = p(1-p)(\frac{1}{n_1} - \frac{1}{n_2})$$

Här är p det gemensamma värdet under H_0, vilket skattas som

$$p^* = \frac{x_1 + x_2}{n_1 + n_2}$$

Vi bilar då teststorheten

$$T = \frac{p_1^* - p_2^*}{\sqrt{p^*(1-p^*)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

som under H_0 är approximativt $\mathsf{N}(0,1)$ fördelad.

Teckentest för medianen

- Medianen för en kontinuerlig fördelning är definierad som det tal M så att $\mathrm{P}(X < M) = \mathrm{P}(X > M) = 1/2$.
- Givet ett stickprov av storlek n vill vi testa $H_0 : M = M_0$ mot en ensidig eller tvåsidig mothypotes.
- Låt Q_+ vara antalet observerade värden större än M.
- Låt Q_- vara antalet observerade värden mindre än M.
- Under H_0 är $Q_+ \sim \mathrm{Bin}(n, 1/2)$ och $Q_- \sim \mathrm{Bin}(n, 1/2)$.
  - Om $H_1 : M < M_0$, förkasta H_0 om Q_+ är för liten.
  - Om $H_1 : M > M_0$, förkasta H_0 om Q_- är för liten.
  - Om $H_1 : M \neq M_0$, förkasta H_0 om $\min(Q_+, Q_-)$ är för liten.
- Teststorheten är binomialfördelad, vi kan direkt beräkna p-värdet.
- Om $X_i - M = 0$ räknar vi den observationen till den minsta av Q_+ eller Q_- eftersom det inte motsäger H_0.

Wilcoxons ranktest

- Beräkna alla absolutdifferenser $|X_i - M_0|$ och ordna dessa i ökande storleksordning.
- Låt R_i vara ranken för den $i$te absolutdifferensen gånger tecknet på motsvarande avvikelse.
- Beräkna Wilcoxons teststorheter

$$W_+ = \sum_{i:R_i>0} R_i, \qquad |W_-| = \sum_{i:R_i<0} |R_i|$$

- Definiera teststorheten $W = \min(W_+, |W_-|)$. Kritiska värden för W finns tabulerad för olika n och α.
- Om två absolutdifferenser är lika stora så tilldelar vi dem medelvärdet av motsvarande ranker. Till exempel om vi observerar differenser 2 och -2 som skulle få rankerna 4 och 5 så tilldelar vi dem båda rank 4.5 gånger motsvarande tecken.

Exempel för stickprov i par

I fallet med parade observationer $(X_1, Y_1), \ldots, (X_m, Y_m)$ bildar vi först differenserna $X_i - Y_i$ och använder sedan ett ranktest på differenserna för att testa om differensernas median är noll.

Exempel 10.6.2 i boken. Vi vill testa hur mycket minne två statistiska paket använder vid analys av ett dataset. Vi gör mätningar (i Kb).

Program	Paket X	Paket Y	$D_i = X_i - Y_i$
1	512	500	12
2	650	600	50
3	890	890	0
4	410	400	10
5	1050	1025	25
6	1500	1400	100
7	600	625	-25
8	750	710	40

Exempel (forts)

Vi vill testa

$$H_0 : M_X = M_Y$$
$$H_1 : M_X \neq M_Y$$

Vi ordnar differenserna och beräknar rankerna

$\lvert D_i \rvert$:	0	10	12	25	25	40	50	100
Rank:	1	2	3	4.5	4.5	6	7	8
R_i :	-1	2	3	-4.5	4.5	6	7	8

Vi har nu

$$W_+ = 2 + 3 + 4.5 + 6 + 7 + 8 = 30.5$$

och $|W_-| = |-1| + |-4.5| = 5.5$. Eftersom vi har ett tvåsidigt test använder vi $W = \min(|W_-|, W_+) = 5.5$. Från tabell ser vi att den kritiska punkten för ett tvåsidigt test med $\alpha = 0.1$ är 6. Eftersom $W = 5.5 < 6$ så kan vi förkasta H_0.

Wilcoxons ranksummetest

Antag att vi har två oberoende stickprov $X_1, \ldots, X_m$ och $Y_1, \ldots, Y_n$ från några fördelningar X respektive Y och vill jämföra medianerna för de två variablerna.

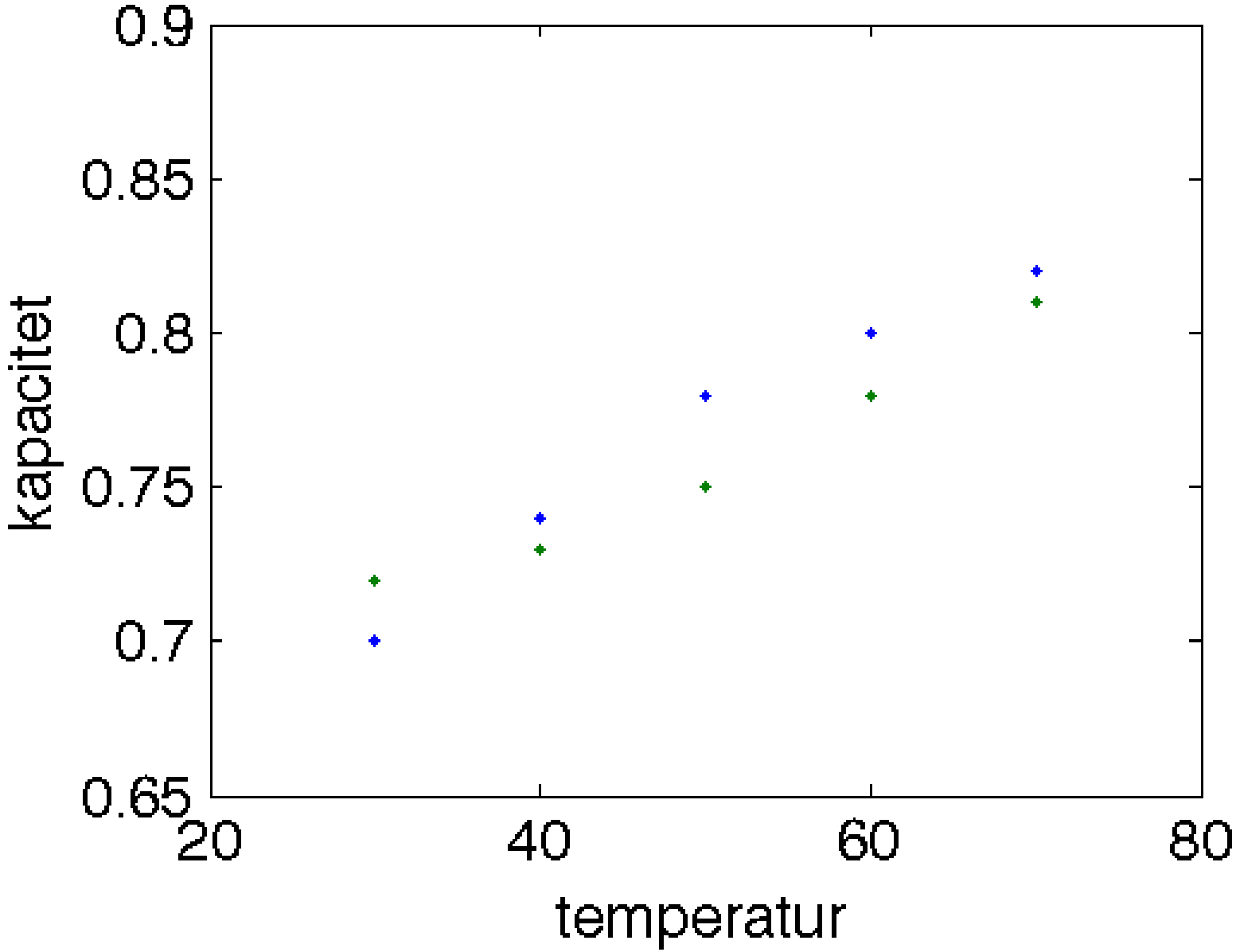
- Ordna de $m + n$ värdena i ökande storleksordning och ge varje observation en rank R_i från 1 till $m + n$.
- Om vi har värden som är lika stora tilldelar vi dem medelranken precis som i rank-testet.
- Vi antar att $m \leq n$ och bildar teststorheten W_m som summan av rankerna associerade med variablerna $X_1, \ldots, X_m$.
- Det kritiska värdet för W_m finns tabulerat för olika värden på m, n, och α.

Exempel, regression

Vi vill undersöka hur ett ämnes specifika värmeskapacitet (ämnets förmåga att magasinera termisk energi) beror av temperaturen.

För var och en av fem temperaturer gör man två mätningar av värmepakaciteten med fäljande resultat:

Temperatur (C)	30	40	50	60	70
värmeskapacitet	0.70	0.74	0.78	0.80	0.82
	0.72	0.73	0.75	0.78	0.81


0.9
0.85
0.8
0.75
0.7
0.65
kapacitet
20
40
60
80
temperatur

Exempel (forts)

Vi ser att värmekapaciteten ser ut att öka för högre temperaturer och vi vill nu anpassa en linje

$$y = \beta_0 + \beta_1 x$$

till datan, så att vi kan prediktera väntevärdet för den specifika värmekapaciteten (y) för en given temperatur (x).

Vi vill också ha ett mått på hur stor de enskilda mätvärdenas spridning kring linjen är. Det är detta vi gör med linjär regression.

Modellbeskrivning

- I exemplet ovan hade vi en *responsvariabel*, y, som är en linjär funktion av en *förklarande variabel* x.
- Förutsättningen för linjär regression är att vi kan välja värdena på x och att dessa kan bestämmas utan fel. Detta gör att vi kan betrakta x-värdena som fixa konstanter.
- För varje värde på x har motsvarande värde på y en viss variation, till exempel på grund av mätfel.
- Vi väljer nu ett antal värden $x_1, \ldots, x_n$ och mäter responsvariabeln för dessa värden. Vi får då mätvärden $y_1, \ldots, y_n$, där y_1 är värdet som är uppmätt då den förklarande variabeln är x_1 och så vidare.

Modellbeskrivning

Vi har alltså talpar (x_i, Y_i) där x_i är ett fixt värde och Y_i är en slumpvariabel. Modellen vi ansätter för Y_i är

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

där ε_i är oberoende $\mathsf{N}(0, \sigma^2)$ slumpvariabeler som beskriver mätfelet och β_0 och β_1 är okända 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.

Exempel (forts)

Vi skattar nu linjen i exemplet ovan.

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = (30 + 30 + 40 + \ldots + 70)/10 = 50$$

$$\bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = (0.70 + 0.72 + \ldots + 0.81)/19 = 0.763$$

$$S_{xx} = \sum_{i=1}^{10} x_i^2 - 10\bar{x}^2 = 27000 - 10 \cdot 50^2 = 2000$$

$$S_{yy} = \sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 = 5.8367 - 10 \cdot 0.763^2 = 0.01501$$

$$S_{xy} = \sum_{i=1}^{10} x_i y_i - 10\bar{x}\bar{y} = 386.8 - 10 \cdot 50 \cdot 0.763 = 5.3$$

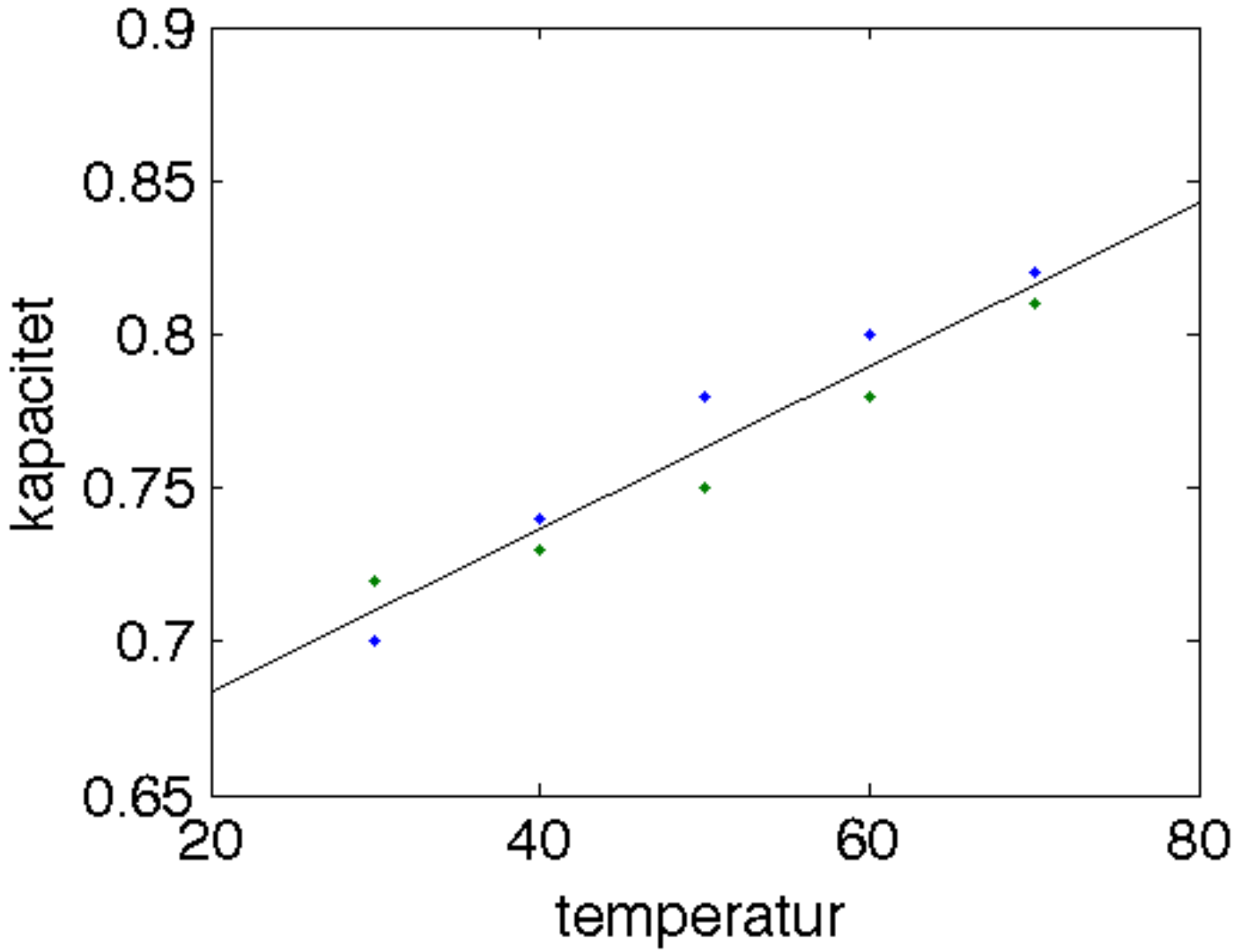
Exempel (forts)

och får därför skattningarna

$$\beta_1^* = S_{xy}/S_{xx} = 5.3/2000 = 0.00265$$
$$\beta_0^* = \bar{y} - \beta_1^* \bar{x} = 0.6305$$
$$s^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 0.00012.$$

En skattning av standardavvikelsen är $s = \sqrt{0.00012} = 0.011$.

Den skattade regressionslinjen $\beta_0 + \beta_1 x$



Egenskaper hos skattarna

Eftersom $\beta_0^* = \bar{y} - \beta_1^* \bar{x}$ så är

$$\bar{y} = \beta_0^* + \beta_1^* \bar{x}$$

vilket betyder att den skattade linjen alltid går genom punkten $(\bar{x}, \bar{y})$.

Detta betyder att om vi centrerar datan, $(x_i - \bar{x}, y_i - \bar{y}), i = 1, \ldots, n$ så har vi att

$$y_i - \bar{y} = \beta_1^*(x_i - \bar{x})$$

och vi har då bara en parameter, β_1, att skatta.

Vi kan skriva den skattade linjens ekvations som antingen

$$y = \beta_0^* + \beta_1^* x$$

eller

$$y = \bar{Y} + \beta_1^*(x - \bar{x})$$

där β_0^*, β_1^* och $\bar{Y}$ är slumpvariabler.

Sats

Vi har att

$$\mathsf{E}(\bar{Y}) = \beta_0 + \beta_1 x \qquad \mathsf{V}(\bar{Y}) = \frac{\sigma^2}{n}$$

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

Vi ser alltså att β_1^* är väntevärdesriktig.

Fördelen med representationen

$$y = \bar{Y} + \beta_1^*(x - \bar{x})$$

är att $\mathsf{C}(\bar{Y}, \beta_1^*) = 0$. Detta göra att vi enkelt kan beräkna variansen för linjen genom att summera osäkerheten i höjdled, $\mathsf{V}(\bar{Y})$, samt osäkerheten i lutning, $\mathsf{V}(\beta_1^*)$.

Sats

Om $\mu_Y^*(x_0) = \beta_0^* + \beta_1^* x_0$ är den skattade linjens värde för ett fixt x_0 så är

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

samt

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

Anmärkning

Om vi i satsen ovan sätter $x_0 = 0$ ser vi att

$$\mathsf{E}(\beta_0^*) = \beta_0$$

samt att

$$\mathsf{V}(\beta_0^*) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

Det framgår också i satsen att osäkerheten i linjenskattningen ökar med x_0:s avstånd till $\bar{x}$.
Vi ser också från de två satserna att β_0^*, β_1^* och $\mu_Y^*(x_0)$ alla är väntevärdesriktiga.

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

Om man är intresserad av var en framtida observation Y kommer ligga för ett visst x_0, och man använder då ett *prediktionsintervall*. Skillnaden mellan ett konfidensintervall för μ_Y och ett prediktionsintervall för Y är att konfidensintervallet talar om var väntevärdet troligen ligger, medan prediktionsintervallet talar om var en framtida observation troligen ligger.
Eftersom vi har en viss spriding kring linjen för de faktiska observationerna så måste prediktionsintervallet vara bredare än konfidensintervallet, och man kan visa att

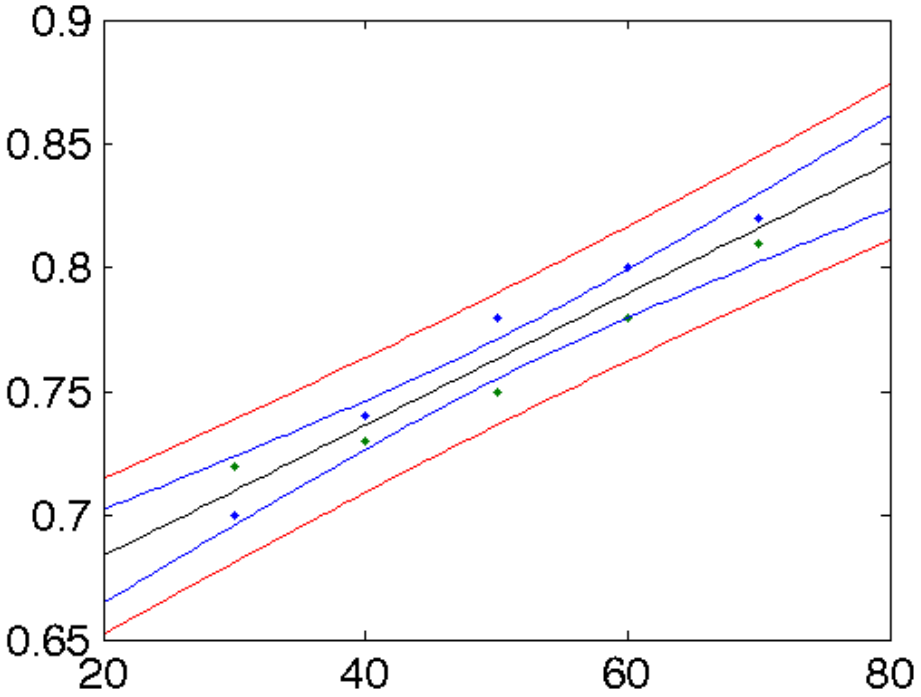
$$Y^*(x_0) \sim \mathsf{N}\left(\beta_0 + \beta_1 x_0, \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})\right).$$

Ett prediktionsintervall bildas nu som

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

Exempel (forts)

Vi beräknar konfidensintervall för regressionslinjen samt prediktionsintervall för framtida observationer för exemplet med värmekapaciteten. De blå linjerna visar övre och undre gränser för konfidensintervallet för alla värden på x och de röda linjerna är gränserna för motsvarande prediktionsintervall.



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.