Föreläsning 12: Linjär regression

Matematisk statistik

David Bolin
Chalmers University of Technology
Oktober 7, 2015

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.

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 $W = \min(|W_-|, W_+) = 5.5$. Enligt tabell är 6 den kritiska punkten med $\alpha = 0.1$: $W = 5.5 < 6$, vi kan 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.

Vi 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 (I)

Vi har två olika typer av värmepannor och vill testa om märke A värmer upp ett rum snabbare än märke B. Vill testa

$$H_0 : M_A = M_B$$
$$H_1 : M_A < M_B$$

Mätningarna är

Märke B:	69.3	56.0	22.1	47.6	53.2	48.1	23.2	13.8
	52.6	34.4	60.2	43.8				
Märke A:	28.6	25.1	26.4	34.9	29.8	28.4	38.5	30.2
	30.6	31.8	41.6	21.1	36.0	37.9	13.9	

Exempel (II)

Vi poolar mätningarna och ordnar dem i stigande storleksordning:

Obs:	13.8	13.9	21.1	22.1	23.2	25.1	26.4	28.4	28.6 . . .
Märke:	B	A	A	B	B	A	A	A	A . . .
Rank:	1	2	3	4	5	6	7	8	9 . . .

Stickprov B är det mindre stickprovet och teststorheten blir därför

$$W_m = 1 + 4 + 5 + \ldots = 212$$

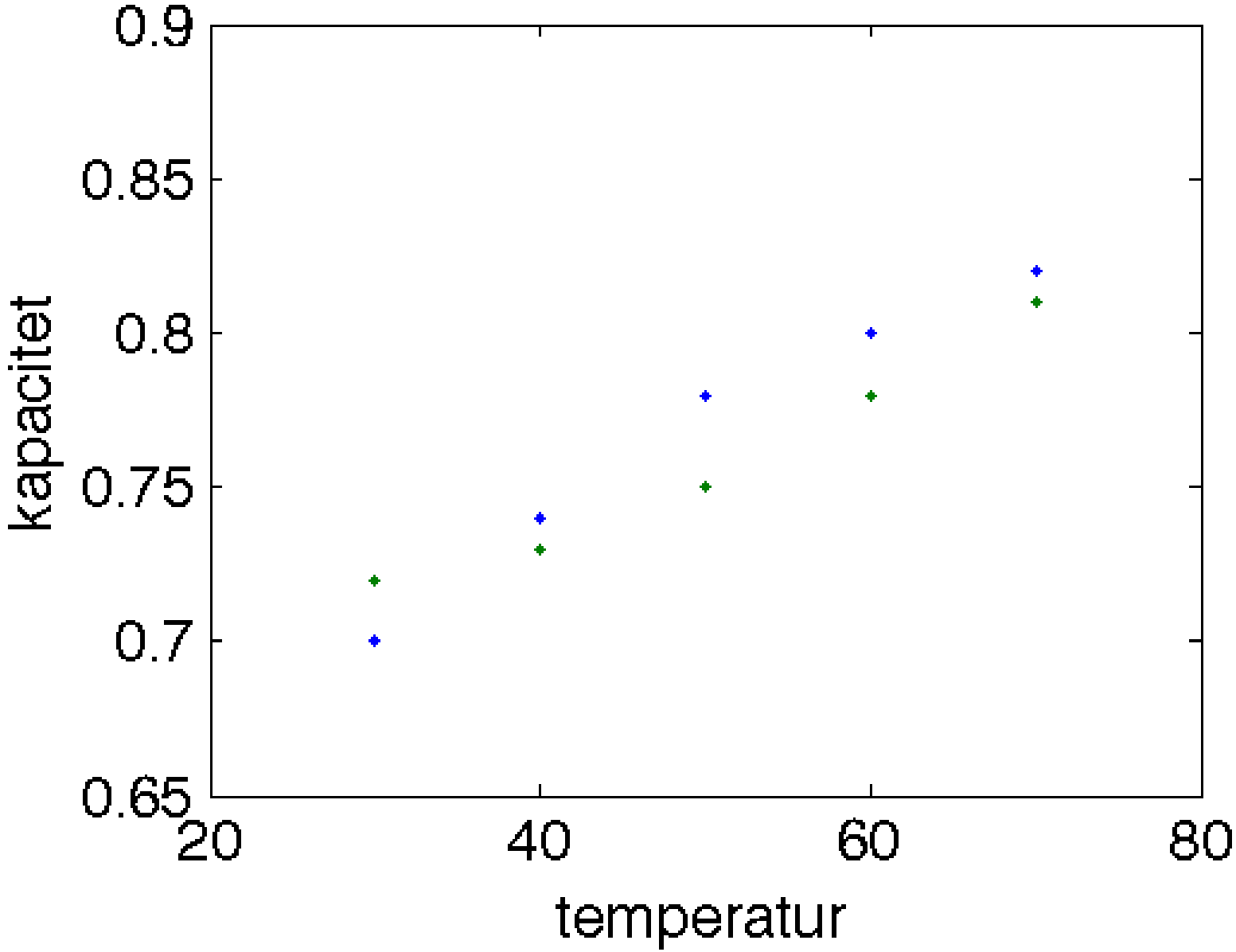
Vi slår upp det kritiska värdet för teststorheten i tabell. Med $m = 12, n = m + 3 = 15$ och $\alpha = 0.05$ är det kritiska värdet 202. Eftersom $W_m = 212 > 202$ kan vi förkasta H_0.

Exempel

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

Skattningar (I)

Vi skattar nu linjen i exemplet ovan.

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

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

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

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

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

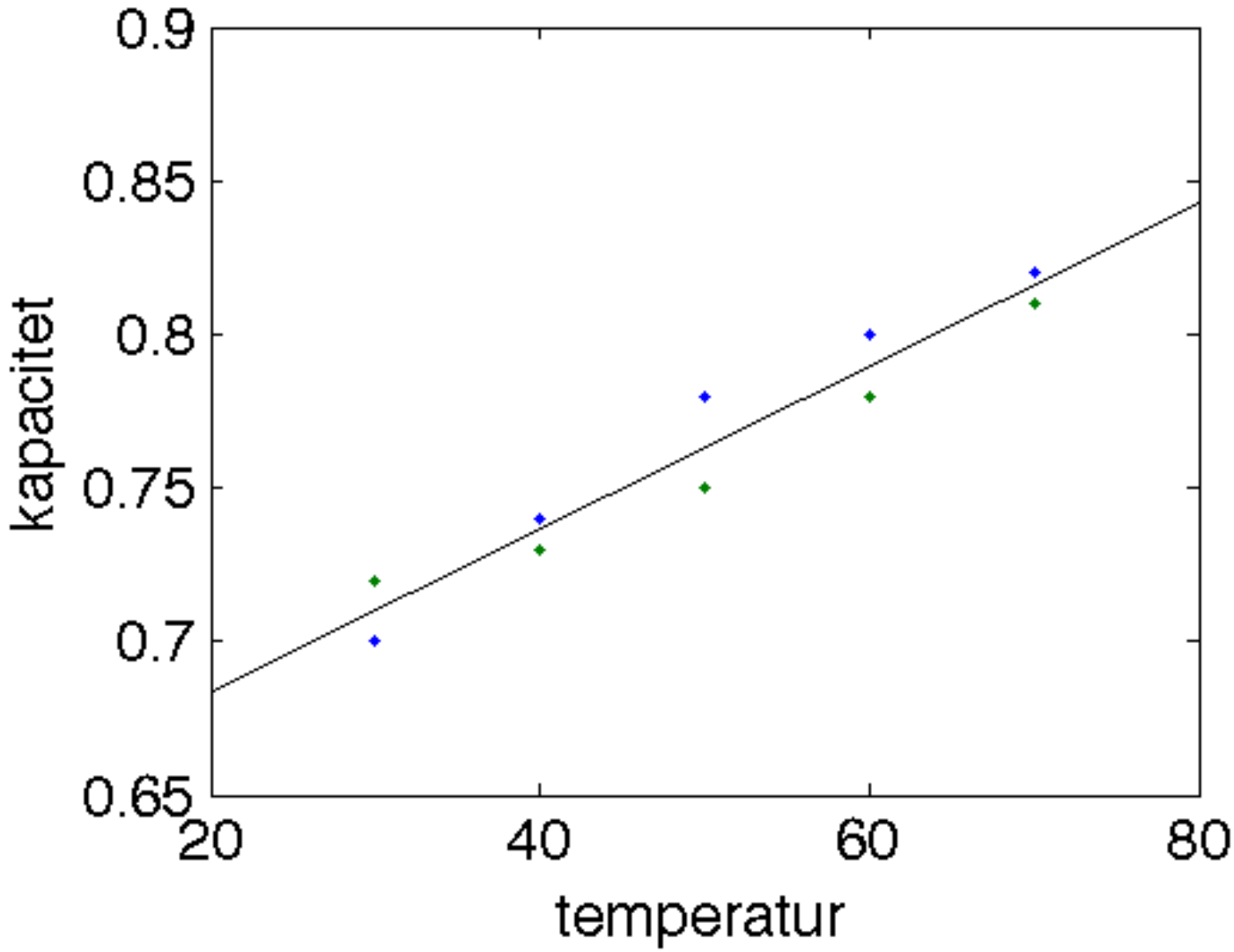
Skattningar (II)

och får därför skattningarna

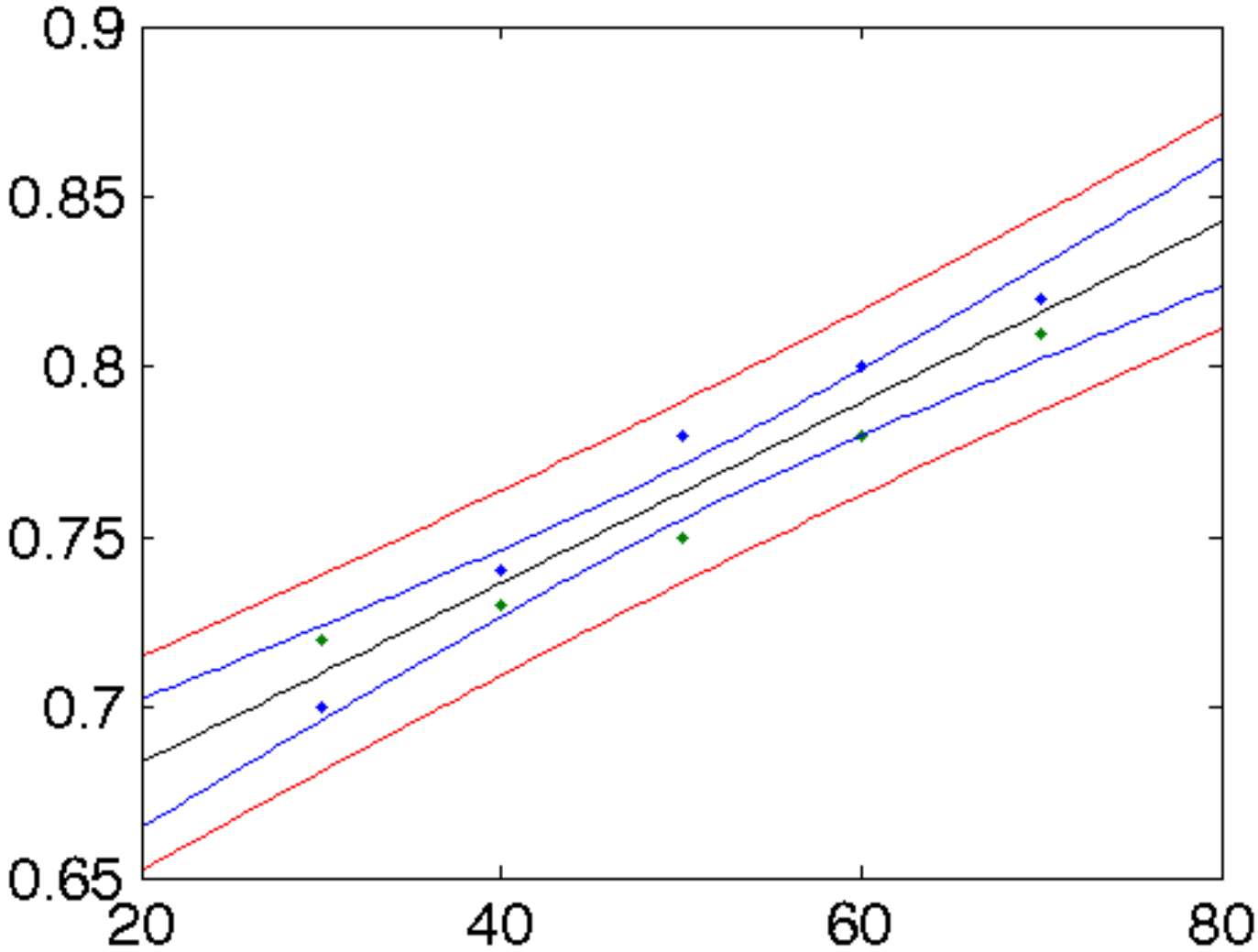
$$\beta_1^* = S_{xy}/S_{xx} = 5.3/2000 = 0.00265$$
$$\beta_0^* = \bar{y} - \beta_1^* \bar{x} = 0.6305$$
$$s^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = 0.00012.$$

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

Den skattade regressionslinjen $\beta_0 + \beta_1 x$



Konfidensintervall och 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.