Aalto-universitetet

MS-A0509 Grundkurs i sannolikhetskalkyl och statistik Gripenberg
Övning 6
Vecka 7, 9-13.2.2015

Besvara Stack-uppgifterna (`stack3.aalto.fi/course/view.php?id=18`)
senast 16.2.2015 kl. 12.00.

Lämna in lösningarna till I-uppgifterna senast 16.2.2015 kl. 12.00.
Kom ihåg att skriva ditt namn och studentnummer!

I1. I tabellen nedan finns för några år antalet skolor i Finland som gav undervisning på gymansienivå:

År	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Antal	483	479	471	461	449	449	441	439	433	428	421

Vi antar att (åtminstone approximativt) $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$, där Y_j är antalet skoler som gav undervisning på gymnasienivå år x_j och slumpvariablerna ε_j är $\mathrm{N}(0, \sigma^2)$-fördelade och oberoende.

(a) Bestäm estimat b_0 och b_1 för koefficienterna i regressionsmodellen $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ med minsta kvadratmetoden.

(b) Beräkna estimatet s^2 för restvariansen och testa nollhypotesen $\mathrm{H}_0 : \beta_1 \geq -5$ på signifikansnivån 1%.

(c) Vad är modellens förklaringsgrad?

I tabellen nedan finns data som underlättar räkningarna

$\overline{y}$	s_x^2	s_y^2	s_{xy}
450.36	11	430.85	-68

I2. I Danmark var under åren 2004-2013 andelen (i procent) av befolkning som var 65 år eller äldre följande:

År	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Andel	14.9	15.0	15.2	15.3	15.6	15.9	16.3	16.8	17.3	17.8

Vi antar att (åtminstone approximativt) $Y_j = \beta_0 + \beta_1(x_j - 2015) + \varepsilon_j$, där Y_j är andelen år x_j av befolkningen i Danmark som äldre än 65 år och slumpvariablerna ε_j är $\mathrm{N}(0, \sigma^2)$-fördelade och oberoende.

(a) Bestäm estimat b_0 och b_1 för koefficienterna i regressionsmodellen $Y_i = \beta_0 + \beta_1(x_i - 2015) + \epsilon_i$ med minsta kvadratmetoden. Vad är β_0?

(b) Beräkna estimatet s^2 för restvariansen och testa nollhypotesen $\mathrm{H}_0 : \beta_0 = 18.8$ på signifikansnivån 1%

I tabellen nedan finns data som underlättar räkningarna

$\overline{y}$	s_x^2	s_y^2	r_{xy}
16.01	9.1667	1.0188	0.97260

När du byter ut talen x_j mot talen $x_j - 2015$ är det endast medelvärdet $\overline{x}$ som måste räknas om.

I3. Utsläppen av växthusgaser i samband med avfallshantering i Finland var under åren 2001-2010 följande (i miljoner ton CO2-ekv.):

2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
3.14	2.92	2.75	2.61	2.40	2.46	2.38	2.28	2.19	2.19

Bestäm med stöd av dessa siffror ett 95% konfidensintervall för väntevärdet av utsläppen år 2014 om du antar att utsläppen kan beskrivas med modellen $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$ där Y_j är utsläppen år x_j och slumpvariablerna ε_j är oberoende och $\mathrm{N}(0, \sigma^2)$-fördelade.

Kun kan använda följande data:

$\overline{x}$	$\overline{y}$	s_x^2	s_y^2	s_{xy}
2005.5	2.532	9.1667	0.10188	-0.92444

och kom ihåg att om du byter ut årtalen mot talen $x_j - 2014$ så är det endast $\overline{x}$ som förändras och du skall räkna ett konfidensintervall för parametern β_0.

Svar: [1.4501, 1.8995]

I4. Med hjälp av observationerna (x_j, y_j), $j = 1, \ldots, n$ har vi räknat ut koefficienterna i regressionslinjen $y = b_0 + b_1 x$ i det fall att x är den "förklarande" variabeln (dvs. då vi antar att $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$) och också för den "inversa" regressionslinjen $x = a_0 + a_1 y$ där y är den "förklarande" variabeln (dvs. då vi antar att $X_j = \alpha_0 + \alpha_1 y_j + \varepsilon_j$). Vad kan man säga om följande påståenden (då vi antar att $s_x > 0$ och $s_y > 0$):

(a) Linjerna $y = b_0 + b_1 x$ och $x = a_0 + a_1 y$ sammanfaller om och endast om $|r_{xy}| = 1$.

(b) Ifall nollhypotesen $\mathrm{H}_0 : \beta_1 = 0$ förkastas på sigifikansnivån 0.01 så förkastas också nollhypotesen $\mathrm{H}_0 : \alpha_1 = 0$ på signifikansnivån 0.01.

Motivera dina svar!

Ledning: (a) Regressionslinjerna kan skrivas i formen $y - \overline{y} = b_1(x - \overline{x})$ *och* $x - \overline{x} = a_1(y - \overline{y})$. *(b) Kom ihåg hur man kan skriva testvariabeln då man testar* $\beta_1 = 0$ *med hjälp av korrelationskoefficienten och vad blir på motsvarande sätt testvariabeln då man testar nollhypotesen* $\alpha_1 = 0$*?*

I5. Vi har ett observerat stickprov (x_i, y_i) $i = 1, \ldots, 25$ och vi har räknat ut stickprovskorrelationen som blev $r_{xy} = -0.47$. Bilda ett (approximativt) symmetriskt 99% konfidensintervall för korrelationen ρ_{XY} med hjälp av Fischers transformation som säger att $\frac{1}{2}\ln\left(\frac{1+R_{xy}}{1-R_{xy}}\right) \sim_a \mathrm{N}\left(\frac{1}{2}\ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right), \frac{1}{n-3}\right)$ tex. på följande sätt:

(a) Bestäm talet z så att $\Pr(-z \le Z \le z) \approx 0.99$ om $Z \sim_a \mathrm{N}(0,1)$.

(b) Välj $Z = \dfrac{\frac{1}{2}\ln\left(\dfrac{1+r_{xy}}{1-r_{xy}}\right) - \frac{1}{2}\ln\left(\dfrac{1+\rho_{xy}}{1-\rho_{xy}}\right)}{\sqrt{\frac{1}{n-3}}}$ (i enlighet med Fishers transformation) och bestäm tal a och b så att

$$-z \le Z \le z \quad \Leftrightarrow \quad a \le \ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right) \le b.$$

(c) Bestäm talen r_L och r_U så att

$$a \le \ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right) \le b \quad \Leftrightarrow \quad r_L \le \rho_{XY} \le r_U.$$

Konfidensintervallet är nu $[r_L, r_U]$.

Svar: $[-0.785, 0.039]$