Aalto-universitetet

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

**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 $\varepsilon_j$ 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 |

*Lösning:* (a) Medelvärdet av $x$-värdena är $\overline{x} = \frac{1}{11}\sum_{j=2003}^{2013} j = 2008$ så att estimaten för parametrarna $\beta_1$ och $\beta_0$ blir

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{-68}{11} = -6.1818$$
$$b_0 = \overline{y} - b_1\overline{x} = 12863.$$

(b) Korrelationskoefficienten är

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} = \frac{-68}{\sqrt{11 \cdot 430.85}} = -0.98776$$

och då blir estimatet för restvariansen

$$s^2 = \frac{n-1}{n-2} s_y^2 (1 - r_{xy}^2) = \frac{10}{9} \cdot 430.85 \cdot (1 - (-0.98776)^2) = 11.647.$$

När vi testar nollhypotesen $\beta_1 \leq 5$ så använder vi testvariabeln

$$W_1 = \frac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}},$$

som är $\mathrm{t}(n-2)$-fördelad. Denhär testvariabeln får värdet

$$w_1 = \frac{-6.1818 + 5}{\sqrt{\frac{11.647}{10 \cdot 11}}} = -3.6319$$

Eftersom alternativet till nollhypotesen är ensidigt blir $p$-värdet

$$p = F_{\mathrm{t}(9)}(-3.6319) = 0.0027,$$

så vi kan förkasta nollhypotesen på signifikansnivån 0.01.

(c) Förklaringsgraden är enligt definitionen $r_{xy}^2 = 0.97567$.

---

**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 $\varepsilon_j$ 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 $\beta_0$?

(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.

*Lösning:* (a) Om vi byter ut talen $x_j$ mot talen $\tilde{x}_j = x_j - 2015$ så blir $\overline{\tilde{x}} = -6.5$. För att räkna ut $b_1$ kan vi använda det faktum att $r_{xy}$ är givet och att

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} \cdot \sqrt{\frac{s_y^2}{s_x^2}} = r_{xy} \cdot \sqrt{\frac{s_y^2}{s_x^2}} = 0.9726 \cdot \sqrt{\frac{1.0188}{9.1667}} = 0.32424.$$

Estimatet $b_0$ för $\beta_0$ blir

$$b_0 = \overline{y} - b_1\overline{\tilde{x}} = 16.01 - 0.32424 \cdot (-6.5) = 18.118$$

(b) Estimatet för restvariansen är

$$s^2 = \frac{n-1}{n-2} s_y^2(1 - r_{xy}^2) = \frac{9}{8} \cdot 1.0188 \cdot (1 - 0.97260^2) = 0.061949.$$

När vi testar nollhypotesen $\beta_0 \leq 19$ så använder vi testvariabeln

$$W_0 = \frac{B_0 - \beta_0}{\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{\tilde{x}}^2}{(n-1)s_x^2}\right)}},$$

som är $\mathrm{t}(n-2)$-fördelad. Denhär testvariabeln får värdet

$$w_1 = \frac{18.118 - 18.8}{\sqrt{0.061949 \cdot \left(\frac{1}{10} + \frac{(-6.5)^2}{9 \cdot 9.1667}\right)}} = -3.5023.$$

Eftersom alternativet till nollhypotesen är dubbelsidigt blir $p$-värdet

$$p = 2 \cdot F_{\mathrm{t}(8)}(-3.5023) = 0.0081 < 0.01$$

så vi kan förkasta nollhypotesen på signifikansnivån 0.01.

---

**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 $\varepsilon_j$ 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 $\beta_0$.

Svar: $[1.4501,\ 1.8995]$

*Lösning*: Om $\tilde{x}_j = x_j - 2014$ så är $\overline{\tilde{x}} = -8.5$ men $s_{\tilde{x}}^2 = s_x^2$ ja $s_{\tilde{x}y} = s_{xy}$.

Vi räknar först ut estimaten för parametrarna $\beta_1$ och $\beta_0$ och de blir

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{-0.92444}{9.1667} = -0.10085,$$
$$b_0 = \overline{y} - b_1\overline{\tilde{x}} = 2.532 - (-0.10085)\cdot(-8.5) = 1.6748.$$

För att kunna räkna ett konfidensintervall behöver vi estimatet $s^2$ för restvariansen och först skall vi räkna ut stickprovskorrelationen som är

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} = \frac{-0.92444}{\sqrt{9.1667 \cdot 0.10188}} = -0.95659.$$

Estimatet för stickprovsvariansen är

$$s^2 = \frac{n-1}{n-2} s_y^2 (1 - r_{xy}^2) = \frac{9}{8} \cdot 0.10188 \cdot (1 - (-0.95659)^2) = 0.0097349.$$

Testvariabeln

$$W_0 = \frac{B_0 - \beta_0}{\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{\tilde{x}}^2}{(n-1)s_x^2}\right)}},$$

är $\mathrm{t}(n-2)$-fördelad så att

$$\Pr(-2.306 \le W_0 \le 2.306) = 0.95$$

eftersom $2.306 = F_{\mathrm{t}(8)}^{-1}(0.975)$ av vilket följer att

$$\Pr\left(B_0 - 2.306\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{\tilde{x}}^2}{(n-1)s_x^2}\right)} \le \beta_0 \le B_0 + 2.306\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{\tilde{x}}^2}{(n-1)s_x^2}\right)}\right) = 0.95.$$

Konfidensintervallet är alltså

$$\left[B_0 - 2.306\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{\tilde{x}}^2}{(n-1)s_x^2}\right)},\ B_0 + 2.306\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{\tilde{x}}^2}{(n-1)s_x^2}\right)}\right],$$

och om vi nu sätter in de värden vi har för de olika variablerna så får vi intervallet

$$\big[1.4501, 1.8995\big].$$

---

**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 $\alpha_1 = 0$?*

*Lösning:* (a) Båda linjerna går genom punkten $(\overline{x}, \overline{y})$ vilket betyder att linjerna sammanfaller om och endast om de har samma vinkelkoefficient. Nu är $b_1 = r_{xy}\frac{s_y}{s_x}$ och $a_1 = r_{xy}\frac{s_x}{s_y}$ och vinkelkoefficenterna är $b_1$ och $\frac{1}{a_1}$ så att dessa är lika då $b_1 a_1 = 1$, dvs. då $r_{xy}^2 = 1$ eller $r_{xy} = \pm 1$. Påståendet gäller.

(b) Testvariabeln kan i båda fallen eftersom vi testar hypoteserna $\beta_1 = 0$ och $\alpha_1 = 0$ skrivas i formen $\frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$ så det är frågan om samma test och därför gäller även detta påstående.

---

**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 $\rho_{XY}$ 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 $[r_L, r_U]$.

*Lösning:* (a) Eftersom $0.005 + 0.005 = 0.01 = 1 - 0.99$ så är $z = z_{0.005} = F^{-1}_{\mathrm{N}(0,1)}(0.995) = 2.5758$.

Svar: $[-0.785, 0.039]$

(b) Eftersom $r_{xy} = -0.47$ så är $\frac{1}{2}\ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right) = -0.51007$ och $\sqrt{\frac{1}{25-3}} = 0.2132$ så att

$$-z \le Z \le z \quad \Leftrightarrow \quad -2.5758 \le \frac{1}{0.2132}\cdot\left(-0.51007 - \tfrac{1}{2}\ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right)\right) \le 2.5758$$

$$\Leftrightarrow \quad 0.2132\cdot(-2.5758) + 0.51007 \le -\tfrac{1}{2}\ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right) \le 0.2132\cdot 2.5758 + 0.51007$$

$$\Leftrightarrow \quad -2\cdot(0.2132\cdot 2.5758 + 0.51007) \le \ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right) \le 2\cdot(0.2132\cdot(-2.5758) + 0.51007),$$

så att $a = -2\cdot(0.2132\cdot 2.5758 + 0.51007) = -2.1185$ och $b = -2\cdot(0.2132\cdot(-2.5758) + 0.51007) = 0.078181$.

(c) Vi har

$$-2.1185 \le \ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right) \le 0.078181 \quad \Leftrightarrow \quad \mathrm{e}^{-2.1185} \le \frac{1+\rho_{xy}}{1-\rho_{xy}} \le \mathrm{e}^{0.078181}$$

$$\Leftrightarrow \quad \mathrm{e}^{-2.1185}\cdot(1-\rho_{xy}) \le 1+\rho_{XY} \le \mathrm{e}^{0.078181}\cdot(1-\rho_{xy})$$

$$\Leftrightarrow \quad \frac{\mathrm{e}^{-2.1185}-1}{\mathrm{e}^{-2.1185}+1} \le \rho_{XY} \le \frac{\mathrm{e}^{0.078181}-1}{\mathrm{e}^{0.078181}+1}$$

så att

$$r_L = \frac{\mathrm{e}^{-2.1185}-1}{\mathrm{e}^{-2.1185}-1} = -0.78538,$$

$$r_U = \frac{\mathrm{e}^{0.078181}-1}{\mathrm{e}^{0.078181}+1} = 0.039071.$$

---