F16, Enkel linjär regression forts..

Christian Tallberg

Statistiska institutionen

Stockholms universitet

Korrelationskoefficienten

Vi har två stokastiska variabler X och Y. Kovariansen, som är ett mått på hur starkt det linjära sambandet är mellan X och Y, ges av

$$Cov\left(X,Y\right)=E\left(X-\mu_x\right)\left(Y-\mu_y\right)$$

Kovariansen är ett icke-normerat mått på det linjära sambandet vilket gör den svår att tolka. Ett normerat mått på hur starkt det linjära sambandet är mellan X och Y ges av korrelationskoefficienten

$$\rho=\frac{Cov\left(X,Y\right)}{sqrt\left(V\left(X\right)V\left(Y\right)\right)},$$

som alltid ligger i intervallet

$$-1\leq\rho\leq 1.$$

I en inferenssituation då vi har data skattas ρ med

$$\begin{aligned} r &= \frac{\sum\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{sqrt\left(\sum\left(x_i-\bar{x}\right)^2\sum\left(y_i-\bar{y}\right)^2\right)} \\ &= \frac{\sum xy-\frac{\sum x\sum y}{n}}{sqrt\left(\left[\sum x^2-\frac{\left(\sum x\right)^2}{n}\right]\left[\sum y^2-\frac{\left(\sum y\right)^2}{n}\right]\right)} \\ &= b\frac{sqrt\left(\sum x^2-\frac{\left(\sum x\right)^2}{n}\right)}{sqrt\left(\sum y^2-\frac{\left(\sum y\right)^2}{n}\right)}, \end{aligned}$$

som också alltid ligger i intervallet

$$-1\leq r\leq 1.$$

Tecknet på r säger om sambandet är positivt eller negativt (jfr med tecknet på b).

I oljeförbrukningsexemplet blir den skattade korrelationen

$$\begin{aligned} r &= \frac{\sum xy-\frac{\sum x\sum y}{n}}{sqrt\left(\left[\sum x^2-\frac{\left(\sum x\right)^2}{n}\right]\left[\sum y^2-\frac{\left(\sum y\right)^2}{n}\right]\right)} \\ &= \frac{8990.5-\frac{63.2\cdot 3320}{10}}{sqrt\left(\left[874.24-\frac{63.2^2}{10}\right]\left[1412000-\frac{3320^2}{10}\right]\right)} \\ &= -0.989\approx -0.99 \end{aligned}$$

OBS! Tecknet på r anger om lutningen är positiv eller negativ men inte hur kraftig lutningen är (detta anger värdet på regressionskoefficienten b).

Det gäller också att

$$\begin{aligned} r &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{sqrt\left(\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2\right)} \\ &= \frac{sqrt\left(\sum (x_i - \bar{x})(y_i - \bar{y})\right) sqrt\left(\sum (x_i - \bar{x})(y_i - \bar{y})\right)}{sqrt\left(\sum (x_i - \bar{x})^2\right) sqrt\left(\sum (y_i - \bar{y})^2\right)} \\ &= \frac{sqrt(b)\, sqrt\left(\sum (x_i - \bar{x})(y_i - \bar{y})\right)}{sqrt\left(\sum (y_i - \bar{y})^2\right)} \\ &= sqrt\left(\frac{SSR}{SST}\right) = sqrt\left(R^2\right) \end{aligned}$$

I oljeförbrukningsexemplet är

$$\begin{aligned} r &= sqrt\left(R^2\right) = sqrt\left(\frac{SSR}{SST}\right) = sqrt\left(\frac{302866}{309760}\right) \\ &= \pm 0.99 = -0.99 \end{aligned}$$

Den linjära regressionsmodellen

Enkel linjär regression

Vi har *en* förklarande variabel och *den skattade linjen i stickprovet* (observerat medelvärde betingat på x_i) skrivs som

$$\hat{y}_i = a + bx_i.$$

Den skattade modellen (en enskild observation betingat på x_i) ges av

$$\begin{aligned} y_i &= a + bx_i + e_i \\ &= \hat{y}_i + e_i \end{aligned}$$

där

$$e_i = y_i - \hat{y}_i$$

är det observerade värdets avvikelse till linjen (*residual*) för ett givet x_i- värde.

Ekvationen för *regressionslinjen i populationen* (väntevärde betingat på x_i) skrivs som

$$\mu_{y|x} = \alpha + \beta x_i,$$

där α och β nu är parametrar. Regressionsmodellen (en enskild observation betingat på x_i) skrivs som

$$\begin{aligned} Y_i &= \alpha + \beta x_i + \varepsilon_i \\ &= \mu_{y|x} + \varepsilon_i, \end{aligned}$$

där

$$\varepsilon_i = Y_i - \mu_{y|x}$$

är avvikelsen mellan Y_i och det betingade väntevärdet (*felterm*).

Både Y_i och ε_i är stokastiska variabler med väntevärden

$$\begin{aligned} E(Y_i) &= \mu_{y|x} \\ E(\varepsilon_i) &= E\left(Y_i - \mu_{y|x}\right) = E(Y_i) - \mu_{y|x} \\ &= \mu_{y|x} - \mu_{y|x} = 0 \end{aligned}$$

och varians

$$\begin{aligned} V(\varepsilon_i) &= \sigma_\varepsilon^2 \\ V(Y_i) &= V\left(\mu_{y|x} + \varepsilon_i\right) = V(\varepsilon_i) = \sigma_\varepsilon^2. \end{aligned}$$

σ_ε^2 skattas väntevärdesriktigt med s_e^2 i stickprovet.

Skattning av α och β

Notera att väntevärdet i regressionen skrivs som

$$\mu_{y|x} = \alpha + \beta x_i.$$

När vi vill skatta $\mu_{y|x}$ med $\hat{y}$, skattar vi α och β med a och b, respektive. Man kan visa att

$$\begin{aligned} E(a) &= \alpha \\ E(b) &= \beta \\ V(a) &= \sigma_a^2 = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right] \\ V(b) &= \sigma_b^2 = \frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{x})^2} \end{aligned}$$

Då σ_ε^2 (nästan) alltid är okänd används skattningen s_e^2 från stickprovet. Den skattade variansen för a och b ges då av

$$\begin{aligned} s_a^2 &= s_e^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right] \\ s_b^2 &= \frac{s_e^2}{\sum (x_i - \bar{x})^2} \end{aligned}$$

	Jämförelse	
Storhet	Utan förkl var	Regression
StokVariab	X	Y (ε)
observerat	x	y (e)
Väntevär.	$E(X) = \mu$	$E(Y) = \mu_{y\|x}$ $(E(\varepsilon) = 0)$
Varians	$V(X) = \sigma^2$	$V(Y) = \sigma_\varepsilon^2$
Skattad var	$\hat{V}(X) = s^2$	$\hat{V}(Y) = s_e^2$
Punktskatt	$\bar{X}$	$\hat{y} = a + bx$
Väntevär.	$E\left(\bar{X}\right) = \mu$	$E(\hat{y}) = \mu_{y\|x}$
Varians	$V\left(\bar{X}\right) = \frac{\sigma^2}{n}$	$V(\hat{y})$ inklud. $V(a), V(b)$
Skattad var	$\hat{V}\left(\bar{X}\right) = \frac{s^2}{n}$	$\hat{V}(\hat{y})$ inklud. s_a^2, s_b^2 (se nec

Hur pass bra är punktskattningarna $y, \hat{y}, a$ och b. För att få en uppfattning om precisionen i skattningarna vill vi göra k.i eller hypotestest. För detta krävs modellantaganden.

- Värdena på variabeln X antas vara fixa. All slumpmässighet finns i Y.
- För varje fixed värde på X antas feltermerna ϵ_i (observationerna Y_i) vara normalfördelade med väntevärde 0 ($\mu_{y|x}$) och varians σ_ε^2.
- Feltermerna ϵ_i (observationerna Y_i) är oberoende sinsemellan.
- Homoscedastisitet. Variansen för feltermerna ϵ_i (observationerna Y_i) , σ_ε^2, är lika för alla X.

Konfidensintervall

Om feltermerna ϵ_i (observationerna Y_i) är normalfördelade följer (av satsen om linjära kombinationer) att även skattningarna av regressionskoefficienterna a och b måste vara normalfördelade. Alltså $a \sim N\left(\alpha; \sigma_a^2\right)$ och $b \sim N\left(\beta; \sigma_b^2\right)$. Ett $(1-\alpha)\,100$ % k.i. för α ges då av

$$\begin{aligned} & a \pm t_{\alpha/2}\,(n-2)\,s_a \\ = \;& a \pm t_{\alpha/2}\,(n-2)\,sqrt\left(s_e^2\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right]\right) \end{aligned}$$

och för β av

$$\begin{aligned} & b \pm t_{\alpha/2}\,(n-2)\,s_b \\ = \;& b \pm t_{\alpha/2}\,(n-2)\,sqrt\left(\frac{s_e^2}{\sum(x_i-\bar{x})^2}\right). \end{aligned}$$

I oljeförbrukningsexemplet är

$$\begin{aligned} s_a &= sqrt\left(862\left[\frac{1}{10}+\frac{6.32^2}{474.82}\right]\right) = 12.60 \\ s_b &= sqrt\left(\frac{862}{474.82}\right) = 1.35 \end{aligned}$$

Ett 95 % k.i. för α ges då av

$$\begin{aligned} & a \pm t_{0.025}\,(n-2)\,s_a \\ = \;& 492 \pm 2.31 \cdot 12.60 = 492 \pm 29.11, \end{aligned}$$

och för β av

$$\begin{aligned} & b \pm t_{0.025}\,(n-2)\,s_b \\ = \;& -25.3 \pm 2.31 \cdot 1.35 = -25.26 \pm 3.12, \end{aligned}$$

Med 95 % tillförlitlighet ligger α i intervallet $(463; 521)$ och β i intervallet $(-28.4; -22.1)$.

Hypotesprövning

Vi vill testa hypotesen

$$H_0 : \beta = \beta_0$$

mot något av alternativen

$$\begin{aligned} H_1 &: \beta \neq \beta_0 \\ H_1 &: \beta < \beta_0 \\ H_1 &: \beta > \beta_0. \end{aligned}$$

Oftast vill man testa om $\beta = 0$, vilket innebär att det inte finns något linjärt samband. På motsvarande sätt kan man testa α.

Från oljeexemplet vill vi testa om vi har linjärt samband eller inte, dvs

$$\begin{aligned} H_0 &: \beta = 0 \\ H_1 &: \beta \neq 0, \end{aligned}$$

för $\alpha = 0.05$. Då $b \sim N\left(\beta; \sigma_b^2\right)$ har vi att testvariabeln

$$t = \frac{b-\beta_0}{s_b} \sim t\,(n-2) \text{ då } H_0 \text{ är sann.}$$

Nollhypotesen förkastas om $|t_{obs}| > 2.31$. Det observerade värdet är

$$t_{obs} = \frac{-25.3-0}{1.35} = -18.7$$

Nollhypotesen förkastas. Ekvivalenta tolkningar: Det kan anses statistiskt påvisat att

- det finns ett linjärt samband mellan temperatur och oljeförbrukning
- temperaturen förklarar en del av variationen i oljeförbrukningen

En alternativ testvariabel är

$$F = \frac{MSR}{MSE} \sim F(1, n-2) \text{ då } H_0 \text{ är sann.}$$

Nollhypotesen förkastas om $F_{obs} > 5.32$. Det observerade testvärdet är

$$F_{obs} = \frac{302866/1}{6894/(10-2)} = 351$$

Samma slutsats som tidigare. Man kan visa att i den här testsituationen är kvadraten av t-värdet $=$ F-värdet.

Test av intercept

$$\begin{aligned} H_0 &: \alpha = 0 \\ H_1 &: \alpha \neq 0, \end{aligned}$$

på signifikansnivån 0.05. Då $a \sim N\left(\alpha; \sigma_a^2\right)$ har vi att testvariabeln

$$t = \frac{a - \alpha_0}{s_a} \sim t(n-2) \text{ då } H_0 \text{ är sann.}$$

Nollhypotesen förkastas om $|t_{obs}| > 2.31$. Det observerade värdet är

$$t_{obs} = \frac{492 - 0}{12.60} = 39.05$$

Nollhypotesen förkastas. Det kan anses statistiskt påvisat att interceptet skall vara med i modellen.