

F17, Enkel linjär regression forts..

Christian Tallberg

Statistiska institutionen

Stockholms universitet

Då Y_i och ϵ_i är normalfördelade givet ett värde på X är även $\hat{y}_i$ normalfördelad givet ett värde på X , dvs $\hat{y}_i|x_0 \sim N\left(\mu_{y|x_0}; \sigma_{\hat{y}|x_0}^2\right)$ (enligt Satsen om Linjära Kombinationer). Ett $(1 - \alpha)$ 100 %-igt k.i. för $\mu_{y|x_0}$ ges då av

$$a + bx_0 \pm t_{\alpha/2}(n-2) \text{sqrt}\left(s_{\hat{y}|x_0}^2\right).$$

Skattning med regressionslinjen

Det betingade väntevärdet

Vi vill göra ett konfidensintervall för det betingade väntevärdet för Y givet ett x_0 . Det ges av

$$E(Y|x_0) = \mu_{y|x_0} = \alpha + \beta x_0$$

som har punktskattningen

$$\hat{y}|x_0 = a + bx_0$$

Variansen för punktskattningen ges av

$$\sigma_{\hat{y}|x_0}^2 = \sigma_{\epsilon}^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

som skattas med

$$s_{\hat{y}|x_0}^2 = s_e^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

Notera att variansen varierar för olika värden på x_0 . Den är minst då $x_0 = \bar{x}$.

I exemplet vill vi göra en intervallskattning för genomsnittliga oljeförbrukningen då temperaturen är 3.8 grader C . Punktskattningen blir

$$a + bx_0 = 492 - 25.3 \cdot 3.8 = 396,$$

och dess skattade varians

$$s_{\hat{y}|x_0}^2 = 862 \left[\frac{1}{10} + \frac{(3.8 - 6.32)^2}{474.82} \right] = 97.73$$

Ett 95 %-igt k.i. för $\mu_{y|x_0=3.8}$ ges då av

$$\begin{aligned} & 396 \pm 2.31 \text{sqrt}(97.73) \\ & = 396 \pm 22.8. \end{aligned}$$

Med 95 % tillförlitlighet ligger $\mu_{y|x_0=3.8}$ i intervallet (373.2; 428.8).

Skattning av en enskild observation (prediktion)

Vi vill göra ett konfidensintervall för en enskild observation Y_i (prognosintervall, prediktionsintervall) givet ett x_0 , dvs

$$Y_i | x_0 = \alpha + \beta x_0 + \varepsilon_i$$

Punktskattningen blir densamma som tidigare, dvs

$$\hat{y} | x_0 = a + bx_0.$$

Däremot ges variansen nu av

$$\sigma_{y|x_0}^2 = \sigma_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

och den skattade variansen av

$$s_{y|x_0}^2 = s_e^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right].$$

Variansen för en enskild observation givet x_0 är större än för den skattade regressionslinjen givet x_0 , eftersom vi dels har en osäkerhet i den skattade linjen och dels i avvikelser från observationen till linjen.

Ett $(1 - \alpha)$ 100 %-igt prognosintervall för $Y_i | x_0$ ges då av

$$(a + bx_0) \pm t_{\alpha/2}(n-2) s_{y|x_0}.$$

Oljeförbrukningsexemplet då temperaturen är 3.8 grader C:

Den skattade variansen blir

$$s_{y|x_0}^2 = 862 \left[1 + \frac{1}{10} + \frac{(3.8 - 6.32)^2}{474.82} \right] = 959.73$$

Ett 95 %-igt p.i. för $Y_i | x_0 = 3.8$ ges då av

$$\begin{aligned} & 396 \pm 2.31 \sqrt{959.73} \\ & = 396 \pm 71.6. \end{aligned}$$

Med 95 % tillförlitlighet ligger $Y_i | x_0 = 3.8$ i intervallet (324.4; 467.6).

Hypotesprövning av korrelationskoefficienter

Man vill testa

$$H_0 : \rho = 0$$

mot något av följande alternativ

$$H_1 : \rho \neq 0$$

$$H_1 : \rho < 0$$

$$H_1 : \rho > 0.$$

ρ skattas i stickprovet med r . Under de modellantaganden som angetts tidigare samt antagandet att variabeln X också följer en normalfördelning är testvariabeln

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2) \text{ då } H_0 \text{ är sann.}$$

I oljeförbrukningsexemplet vill vi testa om vi har korrelation eller inte, dvs

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0,$$

på signifikansnivån 0.05. Nollhypotesen förkastas om $|t_{obs}| > 2.31$. Det observerade värdet är

$$t_{obs} = \frac{-0.989 \sqrt{10 - 2}}{\sqrt{1 - (-0.989)^2}} = -18.9$$

Nollhypotesen förkastas. Det kan anses statistiskt påvisat att vi har korrelation mellan X och Y . Testfunktionen gäller enbart då man vill testa om $\rho = 0$. Värdet på t_{obs} för denna testvariabel är lika med t_{obs} för testvariabeln då man testat $H_0 : \beta = 0$ (förutom avrundningsfel). Man kan visa att testvariablerna är ekvivalenta. Har vi korrelation så har vi ju ett linjärt samband.