Serik Sagitov, Chalmers Tekniska Högskola, August 19, 2005

8. Simple linear regression

Relation between two continuous variables

X = explanatory variable, Y = dependent variable

data: n paired observations (x_i, y_i)

Marginal sample variances

$s_x^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$, $s_y^2 = \frac{1}{n-1}\Sigma(y_i - \bar{y})^2$

Sample covariance

$c_{xy} = \frac{1}{(n-1)}\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \frac{n}{(n-1)}(\overline{xy} - \bar{x}\bar{y})$

sample correlation coefficient $r = \dfrac{c_{xy}}{s_x s_y}$

Ex 1: heights of fathers and sons

http://www.scc.ms.unimelb.edu.au/discday/dyk/faso.html

X = father's height, Y = son's height

8.1 Least square method

Random response to a known independent variable value

$Y = \beta_0 + \beta_1 x + \epsilon$

random noise $\epsilon \sim \mathrm{N}(0, \sigma^2)$ independent of x

model parameters: β_0, β_1, σ^2

Regression lines

unknown true line $y = \beta_0 + \beta_1 x$

fitted line $y = b_0 + b_1 x$ found from the data (x_i, y_i)

Responses

observed y_i and predicted $\hat{y}_i = b_0 + b_1 x_i$

Least square method leading to MLEs

find b_0 and b_1 by minimizing SSE $= \Sigma(y_i - \hat{y}_i)^2$

$$\boxed{\text{Least square regression line } y = \bar{y} + r \cdot \frac{s_y}{s_x}(x - \bar{x})}$$

Least square estimates

slope $b_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} = r \cdot \frac{s_y}{s_x}$

intercept $b_0 = \bar{y} - b_1\bar{x}$

In contrast to correlation coefficient r, regression coefficient b_1 is neither symmetric nor scale free

8.2 Variance estimation $\boxed{\text{SST} = \text{SSR} + \text{SSE}}$

Total sum of squares

$$\text{SST} = \Sigma(y_i - \bar{y})^2 = (n-1)s_y^2$$

Regression sum of squares

$$\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2 = (n-1)b_1^2 s_x^2$$

Error sum of squares

$$\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 = (n-1)s_y^2(1 - r^2)$$

$$\boxed{\text{Corrected MLE of } \sigma^2\text{:} \quad s^2 = \frac{\text{SSE}}{n-2} = \frac{n-1}{n-2}\, s_y^2(1 - r^2)}$$

Coefficient of determination $r^2 = \frac{\text{SSR}}{\text{SST}}$

proportion of variation in y_i explained by x_i variation

Ex 1: heights of fathers and sons

Point estimates in inches (1 inch = 2.54 cm)

$\bar{x} = 68$, $s_x = 2.7$, $\bar{y} = 69$, $s_y = 2.7$

Fitted regresion line $y = 35 + 0.5 \cdot x$

$r = b_1 \cdot \frac{s_x}{s_y} = 0.5$, coefficient of determination is 25%

8.3 CI and hypothesis testing

Estimates of β_0 and β_1 are unbiased and consistent

$b_1 \sim \mathrm{N}(\beta_1, \frac{\sigma_1^2}{n-1})$, $\sigma_1^2 = \sigma^2/s_x^2$

$b_0 \sim \mathrm{N}(\beta_0, \frac{\sigma_0^2}{n-1})$, $\sigma_0^2 = \sigma_1^2 \cdot \frac{1}{n}\Sigma x_i^2$

negative covariance $\mathrm{Cov}(b_0, b_1) = -\frac{\sigma^2 \cdot \bar{x}}{(n-1)\cdot s_x^2}$

Estimated standard errors

$s_{b_1} = \frac{s}{s_x\sqrt{n-1}}$, $s_{b_0} = s_{b_1}\sqrt{\frac{1}{n}\Sigma x_i^2}$

Exact 100(1–α)% CI for $\beta_i = b_i \pm t_{\alpha/2,n-2} \times s_{b_i}$

two t-distributions $\frac{b_0-\beta_0}{s_{b_0}} \sim t_{n-2}$, $\frac{b_1-\beta_1}{s_{b_1}} \sim t_{n-2}$

Hypothesis testing

test H_0: $\beta_1 = \beta_{10}$, using test statistic $T = \frac{b_1-\beta_{10}}{s_{b_1}}$

null distribution $T \sim t_{n-2}$

Model utility test H_0: $\beta_1 = 0$ (no relationship) test statistic $T = b_1/s_{b_1}$, null distribution: $T \sim t_{n-2}$

Ex 1: heights of fathers and sons

$s^2 = \frac{n-1}{n-2}s_y^2(1-r^2) = 5.47$, $s = 2.34$

$s_{b_1} = \frac{s}{s_x\sqrt{n-1}} = 0.026$

99% CI for β_1 is

$0.5 \pm 2.58 \cdot 0.026 = 0.5 \pm 0.07$

model utility test: $T = \frac{b_1}{s_{b_1}} = 18.9$, reject H_0

8.4 Prediction interval

New observation of independent variable for a given x_{n+1}

$Y_{n+1} = \beta_0 + \beta_1 \cdot x_{n+1} + \epsilon_{n+1}$

Expected value of the new observation

true mean $\mu_{n+1} = \beta_0 + \beta_1 \cdot x_{n+1}$

estimated mean $\hat{\mu}_{n+1} = b_0 + b_1 \cdot x_{n+1}$

$\text{Var}(\hat{\mu}_{n+1}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n-1} \cdot \frac{(x_{n+1}-\bar{x})^2}{s_x^2}$

Estimated s.e. of $\hat{\mu}_{n+1}$: $s_{n+1} = s\sqrt{\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{(n-1)s_x^2}}$

Exact 100(1–α)% CI for the mean μ_{n+1}

$b_0 + b_1 \cdot x_{n+1} \pm t_{\alpha/2,n-2} \cdot s_{n+1}$

Exact 100(1–α)% prediction interval for Y_{n+1}

$b_0 + b_1 \cdot x_{n+1} \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s_{n+1}^2}$

Two sources of prediction uncertainty

$\text{Var}(Y_{n+1} - \hat{\mu}_{n+1}) = \text{Var}(\hat{\mu}_{n+1}) + \sigma^2$

Ex 2: my son's height

Estimated mean height of my son $\hat{\mu}_{n+1} = 35 + 0.5 \cdot 72 = 71$

estimated s.e. of $\hat{\mu}_{n+1}$: $s_{n+1} = 0.11$

95% CI for the mean height of my son $= 71 \pm 0.22$

95% PI for the height of my son is

71 ± 4.6 or between 169 cm and 192 cm

actual heights 68.9 (175 cm) and 71.6 (182 cm)