W12

Inference U_yx* and prediction of Y

Proposition

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$, where x^* is some fixed value of x. Then

1. The mean value of $\hat{Y}$ is

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \mu_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*$$

Thus $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is an unbiased estimator for $\beta_0 + \beta_1 x^*$ (i.e., for $\mu_{Y \cdot x^*}$).

2. The variance of $\hat{Y}$ is

$$V(\hat{Y}) = \sigma_{\hat{Y}}^2 = \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum x_i^2 - (\sum x_i)^2/n}\right] = \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

and the standard deviation $\sigma_{\hat{Y}}$ is the square root of this expression. The estimated standard deviation of $\hat{\beta}_0 + \hat{\beta}_1 x^*$, denoted by $s_{\hat{Y}}$ or $s_{\hat{\beta}_0 + \hat{\beta}_1 x^*}$, results from replacing σ by its estimate s:

$$s_{\hat{Y}} = s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

3. $\hat{Y}$ has a normal distribution.

Inference U_yx*

The variable

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} = \frac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{S_{\hat{Y}}} \qquad (12.5)$$

has a t distribution with $n - 2$ df.

A $100(1 - \alpha)\%$ **CI for** $\mu_{Y \cdot x^*}$, the expected value of Y when $x = x^*$, is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2, n-2} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{y} \pm t_{\alpha/2,\, n-2} \cdot s_{\hat{Y}} \qquad (12.6)$$

PI: Prediction Interval for future Y

A $100(1 - \alpha)\%$ **PI for a future** Y **observation to be made when** $x = x^*$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2, n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

$$= \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2, n-2} \cdot \sqrt{s^2 + s_{\hat{\beta}_0 + \hat{\beta}_1 x^*}^2} \qquad (12.7)$$

$$= \hat{y} \pm t_{\alpha/2, n-2} \cdot \sqrt{s^2 + s_{\hat{Y}}^2}$$

The Sample Correlation Coefficient r

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

s^2 = S_xx / (n-1)

The sample correlation coefficient for the n pairs $(x_1, y_1), \ldots, (x_n, y_n)$ is

$$r = \frac{S_{xy}}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \qquad (12.8)$$

Properties of r

The most important properties of r are as follows:

1. The value of r does not depend on which of the two variables under study is labeled x and which is labeled y.
2. The value of r is independent of the units in which x and y are measured.
3. $-1 \le r \le 1$
4. $r = 1$ if and only if (iff) all (x_i, y_i) pairs lie on a straight line with positive slope, and $r = -1$ iff all (x_i, y_i) pairs lie on a straight line with negative slope.
5. The square of the sample correlation coefficient gives the value of the coefficient of determination that would result from fitting the simple linear regression model—in symbols, $(r)^2 = r^2$.

Not dependent on unit

Inference about the population correlation coefficient

$$\rho = \rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

$$\mathrm{Cov}(X, Y) = \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) p(x, y) & (X, Y) \text{ discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y)\, dx\, dy & (X, Y) \text{ continuous} \end{cases}$$

$$\hat{\rho} = R = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

The joint probability distribution of (X, Y) is specified by

$$f(x, y) = \frac{1}{2\pi \cdot \sigma_1 \sigma_2 \sqrt{1 - \rho^2}}\, e^{-[((x-\mu_1)/\sigma_1)^2 - 2\rho(x-\mu_1)(y-\mu_2)/\sigma_1\sigma_2 + ((y-\mu_2)/\sigma_2)^2]/[2(1-\rho^2)]}$$

$$-\infty < x < \infty$$

$$-\infty < y < \infty \qquad (12.9)$$

where μ_1 and σ_1 are the mean and standard deviation of X, and μ_2 and σ_2 are the mean and standard deviation of Y; $f(x, y)$ is called the **bivariate normal probability distribution.**

Testing for the Absence of Correlation

When H_0: $\rho = 0$ is true, the test statistic

$$T = \frac{R\sqrt{n - 2}}{\sqrt{1 - R^2}}$$

has a t distribution with $n - 2$ df.

Alternative Hypothesis	Rejection Region for Level α Test
H_a: $\rho > 0$	$t \ge t_{\alpha, n-2}$
H_a: $\rho < 0$	$t \le -t_{\alpha, n-2}$
H_a: $\rho \ne 0$	either $t \ge t_{\alpha/2, n-2}$ or $t \le -t_{\alpha/2, n-2}$

A P-value based on $n - 2$ df can be calculated as described previously.

When $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a sample from a bivariate normal distribution, the rv

$$V = \frac{1}{2}\ln\left(\frac{1 + R}{1 - R}\right) \qquad (12.10)$$

has approximately a normal distribution with mean and variance

$$\mu_V = \frac{1}{2}\ln\left(\frac{1 + \rho}{1 - \rho}\right) \qquad \sigma_V^2 = \frac{1}{n - 3}$$

Other inference concerning rho

The test statistic for testing H_0: $\rho = \rho_0$ is

$$Z = \frac{V - \frac{1}{2}\ln[(1 + \rho_0)/(1 - \rho_0)]}{1/\sqrt{n - 3}}$$

Alternative Hypothesis	Rejection Region for Level α Test
H_a: $\rho > \rho_0$	$z \ge z_\alpha$
H_a: $\rho < \rho_0$	$z \le -z_\alpha$
H_a: $\rho \ne \rho_0$	either $z \ge z_{\alpha/2}$ or $z \le -z_{\alpha/2}$

A P-value can be calculated in the same manner as for previous z tests.

$$\left(v - \frac{z_{\alpha/2}}{\sqrt{n - 3}},\ v + \frac{z_{\alpha/2}}{\sqrt{n - 3}}\right)$$

A $100(1 - \alpha)\%$ confidence interval for ρ is

$$\left(\frac{e^{2c_1} - 1}{e^{2c_1} + 1},\ \frac{e^{2c_2} - 1}{e^{2c_2} + 1}\right)$$

where c_1 and c_2 are the left and right endpoints, respectively, of the interval (12.11).

Correlation

The simple regression model

Relationship of 2 variables

x: independent, predictor, explanatory variable

y: dependent, response variable

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

A linear probabilistic model

ϵ — random deviation, random error term

True regression line $y = \beta_0 + \beta_1 x$

$\mu_{Y \cdot x^*}$ = the expected (or mean) value of Y when x has value x^*

$\sigma^2_{Y \cdot x^*}$ = the variance of Y when x has value x^*

$$\mu_{Y \cdot x^*} = E(\beta_0 + \beta_1 x^* + \epsilon) = \beta_0 + \beta_1 x^* + E(\epsilon) = \beta_0 + \beta_1 x^*$$

$$\sigma^2_{Y \cdot x^*} = V(\beta_0 + \beta_1 x^* + \epsilon) = V(\beta_0 + \beta_1 x^*) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$$

Assumption

- Linear relationship between x and y
- Error is normally distributed
- Mean error is 0
- Constant variance
- Observation are independent

Estimating Model Parameters

Least squares estimates

$$f(b_0, b_1) = \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2$$

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

The estimated regression line or least squares line

$$b_1 = \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$S_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n$$

$$S_{xx} = \sum x_i^2 - (\sum x_i)^2/n$$

Only within data range !

Estimate sigma

Residuals $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_n - \hat{y}_n$

Error Sum of Square: variation unexplained by the model

$$\mathrm{SSE} = \sum(y_i - \hat{y}_i)^2 = \sum[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

$$\hat{\sigma}^2 = s^2 = \frac{\mathrm{SSE}}{n - 2} = \frac{\sum(y_i - \hat{y}_i)^2}{n - 2}$$

$$\mathrm{SSE} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

The coefficient of determination, r^2: What propostion of variation could be explained by the model

Total Sum of Square $\mathrm{SST} = S_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n$

$$r^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

SST = SSR + SSE

$$r^2 = 1 - \mathrm{SSE/SST} = (\mathrm{SST} - \mathrm{SSE})/\mathrm{SST} = \mathrm{SSR/SST}$$

SSR: Regression sum of square. variation explained by the model

Regression effect

Inference about slop, beta_1

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \frac{\sum Y_i - \hat{\beta}_1 \sum x_i}{n} \qquad \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})Y_i}{S_{xx}} = \sum c_i Y_i \quad \text{where } c_i = (x_i - \bar{x})/S_{xx}$$

$$\hat{\sigma}^2 = S^2 = \frac{\sum Y_i^2 - \hat{\beta}_0 \sum Y_i - \hat{\beta}_1 \sum x_i Y_i}{n - 2}$$

$$V(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{xx}} \qquad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - (\sum x_i)^2/n.$$

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}}$$

Estimator of Beta_1

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

df = n-2

CI for beta_1

$$P\left(-t_{\alpha/2, n-2} < \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} < t_{\alpha/2, n-2}\right) = 1 - \alpha$$

A $100(1 - \alpha)\%$ **CI for the slope** β_1 of the true regression line is

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} \cdot s_{\hat{\beta}_1}$$

Test procedure

Null hypothesis: H_0: $\beta_1 = \beta_{10}$

Test statistic value: $t = \dfrac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$

Alternative Hypothesis	Rejection Region for Level α Test
H_a: $\beta_1 > \beta_{10}$	$t \ge t_{\alpha, n-2}$
H_a: $\beta_1 < \beta_{10}$	$t \le -t_{\alpha, n-2}$
H_a: $\beta_1 \ne \beta_{10}$	either $t \ge t_{\alpha/2, n-2}$ or $t \le -t_{\alpha/2, n-2}$

A P-value based on $n - 2$ df can be calculated just as was done previously for t tests in Chapters 8 and 9.

The **model utility test** is the test of H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \ne 0$, in which case the test statistic value is the t **ratio** $t = \hat{\beta}_1 / s_{\hat{\beta}_1}$.

Regression and ANOVA

$t^2 = f$ and $t^2_{\alpha/2, n-2} = F_{\alpha, 1, n-2}$.

Table 12.2 ANOVA Table for Simple Linear Regression

Source of Variation	df	Sum of Squares	Mean Square	f
Regression	1	SSR	SSR	SSR / SSE/(n − 2)
Error	$n - 2$	SSE	$s^2 = \frac{\mathrm{SSE}}{n - 2}$	
Total	$n - 1$	SST		