Simple linear regression: distributional results

Distributional results for SLR

In order to make inference for the regression coefficients, it is necessary to obtain the distributions of the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

The key results is Lemma 1.
Its application requires that $\hat{\beta}_0$ and $\hat{\beta}_1$ be expressed as linear combinations of $y_1, y_2, \ldots, y_n$.

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x}) x_i$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$
$$= \sum_{i=1}^{n} \left[(x_i - \bar{x}) y_i - (x_i - \bar{x}) \bar{y} \right]$$
$$= \sum_{i=1}^{n} (x_i - \bar{x}) y_i - \bar{y} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{=0}$$
$$= \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} \overbrace{\boxed{\frac{1}{S_{xx}}}\boxed{(x_i - \bar{x})}}^{a_i} y_i = \sum_{i=1}^{n} a_i y_i \quad \text{where } a_i = \frac{x_i - \bar{x}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n} \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a_i y_i \bar{x} = \sum_{i=1}^{n} \overbrace{\boxed{\left(\frac{1}{n} - a_i \bar{x}\right)}}^{b_i} y_i = \sum_{i=1}^{n} b_i y_i \quad \text{where } b_i = \frac{1}{n} - a_i \bar{x}$$

Theorem 8

Suppose $Y_1, Y_2, \ldots, Y_n$ are independent with

$$E[Y_i] = \beta_0 + \beta_1 x_i \text{ and } var(Y_i) = \sigma^2$$

Then

1. $E[\hat{\beta}_0] = \beta_0$ and $E[\hat{\beta}_1] = \beta_1$
2. $var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$ and $var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$
3. $cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{S_{xx}}$
4. $E[S_e^2] = \sigma^2$.

Theorem 8 (cont.)

5. If, furthermore, $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ independently for $i = 1, 2, \ldots, n$, then

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}}{S_{xx}}\right)\right),$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right),$$

$$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Proof of Theorem 8

(1) $$E[\hat{\beta}_1] = \sum_{i=1}^{n} a_i E(Y_i) \quad \text{by Lemma 1}$$

$$= \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right)(\beta_0 + \beta_1 x_i)$$

$$= \left(\frac{1}{S_{xx}}\right)\left[\beta_0 \underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{=0} + \beta_1 \underbrace{\sum_{i=1}^{n}(x_i - \bar{x}) x_i}_{S_{xx}}\right]$$

$$= \left(\frac{1}{S_{xx}}\right)(\beta_1 S_{xx})$$

$$= \beta_1$$

$$E[\hat{\beta}_0] = \sum_{i=1}^{n} b_i E(Y_i) \quad \text{by Lemma 1}$$

$$= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}\right)(\beta_0 + \beta_1 x_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \beta_0 - \frac{\beta_0 \bar{x}}{S_{xx}} \underbrace{\sum_{i=1}^{n}(x_i - \bar{x})}_{=0} + \frac{\beta_1}{n} \underbrace{\boxed{\sum_{i=1}^{n} x_i}}_{n\bar{x}} - \frac{\beta_1 \bar{x}}{S_{xx}} \underbrace{\sum_{i=1}^{n}(x_i - \bar{x}) x_i}_{S_{xx}}$$

$$= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x}$$

$$= \beta_0$$

Proof of Theorem 8

② $\mathrm{var}(\hat\beta_1) = \sum_{i=1}^{n} a_i^2 \,\mathrm{var}(Y_i)$ by Lemma 1

$$= \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 \sigma^2$$

$$= \frac{\sigma^2}{S_{xx}^2} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{S_{xx}}$$

$$= \frac{\sigma^2}{S_{xx}}$$

$\mathrm{var}(\hat\beta_0) = \sum_{i=1}^{n} b_i^2 \,\mathrm{var}(Y_i)$ by Lemma 1

$$= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}}\right)^2 \sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} - \frac{2\bar{x}(x_i - \bar{x})}{n S_{xx}} + \frac{\bar{x}^2 (x_i - \bar{x})^2}{S_{xx}^2}\right)$$

$$= \sigma^2 \left(\frac{1}{n} - \frac{2\bar{x}}{n S_{xx}} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{=0} + \frac{\bar{x}^2}{S_{xx}^2} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{S_{xx}}\right)$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Proof of Theorem 8

③ $$\operatorname{cov}(\hat\beta_0, \hat\beta_1) = \operatorname{cov}\left(\sum_{i=1}^{n} b_i Y_i, \sum_{j=1}^{n} a_j Y_j\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} b_i a_j \operatorname{cov}(Y_i, Y_j)$$

$$= \sum_{i=1}^{n} \left[b_i a_i \underbrace{\operatorname{cov}(Y_i, Y_i)}_{\operatorname{var}(Y_i)} + \sum_{j \neq i} b_i a_j \overbrace{\operatorname{cov}(Y_i, Y_j)}^{=0} \right]$$

as Y_i and Y_j are independent for $i \neq j$

$$= \sum_{i=1}^{n} a_i b_i \operatorname{var}(Y_i)$$

$$= \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}} \right) \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}} \right)$$

$$= \frac{\sigma^2}{S_{xx}} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{n} - \frac{\bar{x}(x_i - \bar{x})^2}{S_{xx}} \right)$$

$$= \frac{\sigma^2}{S_{xx}} \left[\frac{1}{n} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{=0} - \frac{\bar{x}}{S_{xx}} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{S_{xx}} \right]$$

$$= \frac{\sigma^2}{S_{xx}} (-\bar{x})$$

$$= -\frac{\sigma^2 \bar{x}}{S_{xx}}$$

Corollary 8

If $Y_1, Y_2, \dots, Y_n$ are independently with $Y_i \sim N\left(\beta_0 + \beta_{1x_i}, \sigma^2\right)$, then

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

$$\boxed{\frac{\hat{\beta}_0 - \beta_0}{s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2},}$$

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1)$$

We estimate σ with Se

$$V = \frac{(n-2)\, Se^2}{\sigma^2} \sim \chi^2_{n-2}$$

$$\boxed{\frac{\hat{\beta}_1 - \beta_1}{s_e / \sqrt{S_{xx}}} \sim t_{n-2}.}$$

$$T = \frac{Z}{\sqrt{\frac{V}{n-2}}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\frac{\cancel{\sigma}}{\sqrt{S_{xx}}}}}{\sqrt{\frac{\cancel{(n-2)}\, Se^2}{\cancel{(n-2)}\, \cancel{\sigma^2}}}} = \frac{\hat{\beta}_1 - \beta_1}{\frac{Se}{\sqrt{S_{xx}}}} \sim t_{n-2}$$

Hypothesis tests for β_0 and β_1

The pivotal quantities given in Corollary 8 can be used to derive confidence interval and test of hypotheses for the two regression coefficients, β_0 and β_1.

In practical applications, interest is usually focused on β_1.

Under the assumption of the linear regression model, the hypothesis $H_0: \beta_1 = 0$ can be tested to determine whether there is a significant linear relationship between x and y.

Hypothesis tests for β_1

Testing the hypothesis $H_0: \beta_1 = \beta_{10}$. vs $H_a: \beta_1 \neq \beta_{10}$

test statistic : $t = \dfrac{\hat{\beta}_1 - \beta_{10}}{S_e / \sqrt{S_{xx}}}$ $\sim t_{n-2}$ under H_0

critical region : reject H_0 if $|t| \geq t_{n-2, \frac{\alpha}{2}}$

p-value : $P(|T| \geq |t|)$ where $T \sim t_{n-2}$

CI : $\hat{\beta}_1 \pm t_{n-2, \frac{\alpha}{2}} \dfrac{S_e}{\sqrt{S_{xx}}}$

Hypothesis tests for β_0

Testing the hypothesis $H_0: \beta_0 = \beta_{00}$. vs $H_a: \beta_0 \neq \beta_{00}$.

test statistic: $t = \dfrac{\hat{\beta}_0 - \beta_{00}}{S_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$ $\sim t_{n-2}$ under H_0

critical region: reject H_0 if $|t| \geq t_{n-2,\frac{\alpha}{2}}$.

P-value: $P(|T| \geq |t|)$ where $T \sim t_{n-2}$

CI: $\hat{\beta}_0 \pm t_{n-2,\frac{\alpha}{2}} \, S_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$

Example 3.1

The recorded temperature (°C) of two cities (x and y) for the last 10 days are given in the table below. We would like to investigate whether there is a relationship between the temperature of these two cities.

x	-1	0	2	-2	5	6	8	11	12	-3
y	-5	-4	2	-7	6	9	13	21	20	-9

a) Use the method of least squares to fit a straight line $y = \beta_0 + \beta_1 x$ to these data points.

b) Construct a 95% confidence interval for β_0.

c) Test the hypothesis $H_0\colon \beta_1 = 0$ versus $H_a\colon \beta \neq 0$ using $\alpha = 0.05$ level of significance.

Example 3.1 Solution

$\bar{x} = 3.8, \quad \bar{y} = 4.6$

$\sum_{i=1}^{n} x_i = 38, \quad \sum_{i=1}^{n} y_i = 46, \quad \sum_{i=1}^{n} x_i y_i = 709, \quad \sum_{i=1}^{n} x_i^2 = 408, \quad \sum_{i=1}^{n} y_i^2 = 1302$

a) $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 534.2$

$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 408 - \frac{38^2}{10} = 263.6$

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n} (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}) \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y}\left(\sum_{i=1}^{n} x_i\right) - \bar{x}\left(\sum_{i=1}^{n} y_i\right) + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}
\end{aligned}$$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{534.2}{263.6} = 2.0266$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 4.6 - (2.0266)\, 3.8 = -3.101$

Example 3.1 Solution

b) $CI = \hat{\beta}_0 \pm t_{n-2,\frac{\alpha}{2}} \, Se \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$

$Se^2 = \frac{1}{n-2}\left(S_{yy} - \hat{\beta}_1^2 S_{xx}\right) = \frac{1}{8}\left(1090.4 - 2.0266^2 \times 263.6\right) = 0.9768$

$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left[\sum_{i=1}^{n} y_i\right]^2 = 1090.4$

$t_{8,\,0.025} = 2.306$ $\qquad$ qt(0.975, 8)

$\therefore \; CI = -3.101 \pm 2.306 \sqrt{0.9768\left(\frac{1}{10} + \frac{3.8^2}{263.6}\right)}$

$\approx (-3.998,\ -2.204)$

Example 3.1 Solution

c) $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

test statistic: $t = \dfrac{\hat{\beta}_1 - 0}{S_e / \sqrt{S_{xx}}}$

$$= \frac{2.0266}{\sqrt{\frac{0.9768}{263.6}}}$$

$$\approx 33.2917$$

critical region: reject H_0 if $|t| \geq t_{8, 0.025} = 2.306$.

As t lies within the critical region, we have sufficient evidence (at the 5% significance level) to reject H_0 that the slope of the regression line is 0.