Simple Linear Regression : Part II

April 21st, 2016.

s_x ... (unbiased) estimate for std dev in x

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

s_y ... (unbiased) estimate for std dev in y

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

} measures of Spread.

$cor_{x,y}$ ... empirical correlation between x and y

$$\Rightarrow \quad b_1 = cor_{x,y} \cdot \frac{s_y}{s_x}$$

Least-squares line passes through $(\bar{x}, \bar{y})$

$$\Rightarrow \quad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

(x_i, y_i) $i = 1 \dots n$ observed pairs

↓

fit the least squares line

$$\hat{y} = b_0 + b_1 \cdot x$$

↓

$$x_i \longmapsto \hat{y}_i = b_0 + b_1 \cdot x_i \neq y_i$$

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

From the def'n of the least squares line:

SE .... sum of residuals :

$$SE = \sum e_i = \sum (y_i - \hat{y}_i) = 0$$

SSE ... sums of squares of errors ; sum of squared residuals

$$\sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

sums of squares due to error

SST... total sum of squares:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1) \cdot \underbrace{s_y^2}$$

Empirical variance of y (ubiased)

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SSM (sum of squares due to model)

$$R^2 = 1 - \frac{SSE}{SST} \quad \text{.... (multiple) } R^2$$

R ... coefficient of determination.

Estimate of σ ?

$$s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$$

$$\hat{y} = b_0 + b_1 \cdot x_i$$

estimated from the data $(x_i; y_i)$

$$s = \sqrt{\frac{SSE}{n-2}}$$

Thm. Let

$$Y = \underbrace{\beta_0 + \beta_1 \cdot x}_{} + \underbrace{\varepsilon}_{}$$

be a simple linear regression model

w/ $\varepsilon \sim N(\text{mean}=0, \text{var}=\sigma^2)$.

Let the errors ε_i associated w/ different observations be independent.

Then: $\left.\begin{array}{l} \cdot\ \hat{\beta}_0 \\ \cdot\ \boxed{\hat{\beta}_1} \end{array}\right\}$ are both normally distributed estimators.

- $\mathbb{E}[\hat{\beta}_1] = \beta_1$
- $\text{Var}[\hat{\beta}_1] = \dfrac{\sigma^2}{S_{xx}}$.

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

Remarks:

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

↑ the unbiased estimate of the variance in x

Rewrite $\hat{\beta}_1$ in std units:

$$Z_1 = \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0,1).$$

<u>However</u>: $\frac{SSE}{\sigma^2}$ is $\chi^2(df = n-2)$ and independent from $\hat{\beta}_1$ (and Z_1)

$$\Rightarrow \quad t_{\beta_1} := \frac{Z_1}{\sqrt{\frac{SSE/\sigma^2}{n-2}}} \sim t(df = n-2)$$

$$t_{\beta_1} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}}}{\sqrt{\frac{SSE}{n-2} \cdot \frac{1}{\sigma^2}}} = \frac{\hat{\beta}_1 - \beta_1}{\frac{S}{\sqrt{S_{xx}}}}$$

$$\hat{\beta}_1 = b_1 = cor_{x,y} \frac{S_y}{S_x} = \frac{S_{xy}}{S_{xx}} \quad ; \quad \hat{\beta}_0 = b_0 = \bar{y} - \bar{x} \cdot \hat{\beta}_1$$

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

$\Rightarrow \hat{y} = b_0 + b_1 \cdot x$

$\Rightarrow$ SSE $\Rightarrow$ (A)

Rewrite $\hat{\beta}_1$ in standard units:

$$Z_1 := \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0,1)$$

However: $s^2 = \frac{SSE}{\sigma^2}$ is $\chi^2(df = n-2)$

and independent from $\hat{\beta}_1$ (and Z_1)

$\Rightarrow$

$$t_{\beta_1} := \frac{Z_1}{\frac{\sqrt{SSE}}{\sigma\sqrt{n-2}}} \sim t(df = n-2)$$

$$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{\frac{s}{\sqrt{S_{xx}}}} \sim t(df = n-2)$$

An $(1-\alpha)$ confidence interval for β_1:

$$b_1 \pm t^*_{\alpha/2}(n-2) \cdot \frac{s}{\sqrt{S_{xx}}}$$

Problem. From a data set:

$$\begin{cases} \bar{x} = 3.8\,, \quad \bar{y} = 4.6\,; \\ S_{xx} = 263.6\,; \quad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 534.2. \end{cases}$$

$\hat{\beta}_1$ observation/estimate:

$$b_1 = cor_{x,y} \cdot \frac{s_y}{s_x} = \frac{S_{xy}}{S_{xx}} = \frac{534.2}{263.6} \approx 2.0266;$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 4.6 - 2.0266 \cdot 3.8 = -3.1011$$

An $(1-\alpha)$-confidence interval for β_1 is:

$$b_1 \pm t^*_{\alpha/2}(n-2) \times \frac{s}{\sqrt{S_{xx}}}$$

Problem: From a data set:

$S_{xx} = 263.6$; $S_{xy} = 534.2$;

$\bar{y} = 4.6$; $\bar{x} = +3.8$.

$\hat{\beta}_1$ value: $b_1 = \frac{S_{xy}}{S_{xx}} = \frac{534.2}{263.6} = 2.0266$;

$b_0 = \bar{y} - b_1 \cdot \bar{x} = 4.6 - 2.0266 \cdot 3.8$

$b_0 = -3.1011$

Also: we're given: $(n = 10)$

$\sum_{i=1}^{n} y_i^2 = 1,302$; $S_{yy} = \sum_i (y_i - \bar{y})^2 =$

$= \sum_i y_i^2 - n(\bar{y})^2 = 1090.4$

We want to use:

$(s = ?)$ $\quad s^2 = \frac{SSE}{n-2}$

A teeny bit of algebra:

$SSE = S_{yy} - b_1 \cdot S_{xy} =$

$= 1090.4 - 2.0266 \cdot 534.2 = 7.79$

$s = \sqrt{\frac{7.79}{8}} = \sqrt{0.9738}$

Construct a 95% confidence interval for β_1:

$$t^*_{0.025}(df = 8) = 2.306$$

$\Rightarrow$ margin of error:

$$2.306 \cdot \sqrt{\frac{0.9738}{263.6}} \approx 0.1401$$

$\Rightarrow$ the conf. interval is:

$$2.0266 \pm 0.1401$$

$$= (1.886, 2.1667)$$