Simple Linear Regression Formulas

Simple linear regression model: \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \)

Bivariate data: \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\)

Least squares estimates: slope estimate \( b_1 = r \frac{s_y}{s_x} \), intercept estimate \( b_0 = \bar{y} - b_1 \bar{x} \)

where \( \bar{x} = \frac{\sum x_i}{n}, \bar{y} = \frac{\sum y_i}{n}, s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}, s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}, r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \)

Least squares prediction line: \( \hat{y} = b_0 + b_1 x \)

Statistical Inference: \( \epsilon_i \sim iid N(0, \sigma^2) \) for all \( x \), so \( \mu_y = E(y) = \beta_0 + \beta_1 x \)

The unknown parameter values to be estimated from the data are \( \beta_0, \beta_1, \) and \( \sigma^2 \).

Estimates of \( \beta_0 \) and \( \beta_1 \) are the least squares estimates \( b_0 \) and \( b_1 \), respectively, given above.

NOTE: in the formulas immediately below, \( SSE \) stands for the Sum of the Squares of the Errors. Recall that the error (also called the residual) is defined as \( y_i - \hat{y}_i \), the difference between an observed value \( y_i \) and the corresponding predicted value \( \hat{y}_i \) predicted by the least squares line. \( SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \)

Estimate of \( \sigma^2: \)
\[
\hat{\sigma}^2 = \frac{SSE}{n-2} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}
\]

Alternate formula for \( SSE \)
\[
SSE = (\sum_{i=1}^{n} y_i^2) - b_0 \cdot (\sum_{i=1}^{n} y_i) - b_1 \cdot (\sum_{i=1}^{n} x_i y_i)
\]
Confidence interval for the slope $\beta_1$:

$$b_1 \pm t_{n-2}^* \text{SE}(b_1) \text{ where } SE(b_1) = \frac{s_e}{\sqrt{n-1}s_x}$$

Notation heads-up: $SE(b_1)$ is also sometimes denoted $s_h$

Hypothesis test for the slope $\beta_1$:

$$H_0: \beta_1 = 0; \text{ Test statistic: } t = \frac{b_1 - 0}{SE(b_1)}$$

Predicted value of $y$ for $x = x_v$: \( \hat{y}_v = b_0 + b_1x_v \)

Confidence interval for $\mu_y$ for $x = x_v$:

$$\hat{y}_v \pm t_{n-2}^* SE(\hat{\mu}_v) \quad \text{ where } SE(\hat{\mu}_v) = \sqrt{\left(SE(b_1)\right)^2 \cdot (x_v - \bar{x})^2 + \frac{s_e^2}{n}}$$

Prediction interval for $y_v$:

$$\hat{y}_v \pm t_{n-2}^* SE(\hat{y}_v) \quad \text{ where } SE(\hat{y}_v) = \sqrt{\left(SE(b_1)\right)^2 \cdot (x_v - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$$

Because of the factor $(x_v - \bar{x})^2$ in the above formulas for $SE(\hat{\mu}_v)$ and $SE(\hat{y}_v)$, the confidence intervals for $\mu_y$ and the prediction intervals for $y_v$ for a particular value $x_v$ of $x$ are wider for values of $x_v$ farther from the mean 2.9 of the $x$ data.