Summary for simple linear regression

Dependent data $x_i, y_i \quad i = 1,\ldots,n$

Linear model

$$\hat{y}_i = b_0 + b_1 x_i \quad i = 1,\ldots,n \qquad\qquad y_i = \hat{y}_i + \varepsilon_i$$

$$n\overline{()} = \sum_{i=1}^{n} ()_i; \qquad \rho = \frac{\overline{x'y'}}{\sqrt{\overline{x'^2}\,\overline{y'^2}}}$$

$$b_0 = \overline{y} - b_1\overline{x}\,; \qquad\qquad b_1 = \frac{\overline{x'y'}}{\overline{x'^2}} = \rho\frac{s_y}{s_x}$$

$SST = \sum_{i=1}^{n} y_i'^2 = n\overline{y_i'^2}$ is the total variance of y (with *n-1* d.o.f.)

$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = SST\left(1 - \rho^2\right)$ is the residual (error) variance, with *n-2* d.o.f, or in the case of multiple regression, with K predictors, with *n-K-1* d.o.f.

$SSR = SST - SSE = SST\rho^2$ is the "explained variance", with 1 d.o.f. (or K d.o.f. in the case of multiple regression).

Generalized coefficient of determination R^2:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = \frac{\sum_{i=1}^{n} y_i'^2 - \sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} y_i'^2}$$

: "explained variance" (for the dependent sample)

Estimation of errors in the regression coefficients:

The coefficient b_1, divided by the standard deviation obtained from the sample has a t_{n-2} random distribution:

$$\frac{b_1 - E(b_1)}{\sqrt{\left(\frac{SSE}{n-2}\right) / \sum_{i=1}^{n} x'^2_i}} \sim t_{n-2}$$

This means that we can test whether b_1 is significantly different from zero by using the t table 2 with n-2 d.o.f. at a level of significance of, let's say 5%. The 95% confidence interval for b_1 is

$$\left[b_1 - \sqrt{\left(\frac{SSE}{n-2}\right) / \sum_{i=1}^{n} x'^2_i} * t_{.025,n-2}, b_1 + \sqrt{\left(\frac{SSE}{n-2}\right) / \sum_{i=1}^{n} x'^2_i} * t_{.025,n-2} \right]$$

Similarly

$$\frac{b_0 - E(b_0)}{\sqrt{\left(\frac{SSE}{n-2}\right) * \frac{1}{n} \sum_{i=1}^{n} x'^2_i}} \sim t_{n-2}$$ and the limits of confidence for b_0 are similarly determined.

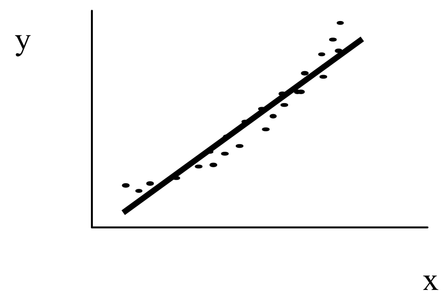
For a new predictor x_0, the limits of confidence of the new prediction can be obtained from the fact that it has a distribution

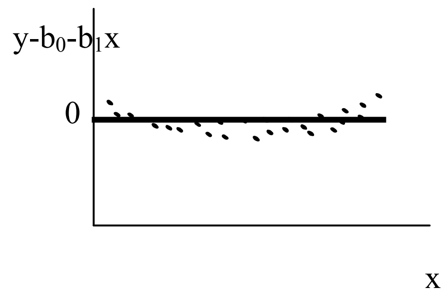
$$\frac{y(x_0) - b_0 - b_1 x_0}{\sqrt{\left(\frac{SSE}{n-2}\right) * \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} x'^2_i}\right)}} \sim t_{n-2}$$

We will see more about this after the multiple regression discussion.

Note that a "naïve" estimation of the forecast error variance $\overline{\varepsilon^2} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \approx \frac{SSE}{n}$ seriously underestimates the error even for the dependent sample, for which the correct formula is $s_\varepsilon^2 = \overline{\varepsilon^2} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{SSE}{n-K-1}$.

Analysis of residuals: Ideally the residuals (forecast errors) should look random, without trends. If we find, for example





If there is a trend, it is better to change variables, for the predictor, e.g., $y = b_0 + b_1 f(x)$, $y = b_0 + b_1 x + b_2 x^2 + ... + b_K x^K$. Note that this is still a linear regression, with multiple predictors, even if the predictors are nonlinear functions of x.