VARIANCE OF RESIDUALS IN SIMPLE LINEAR REGRESSION

ALLEN BACK

Suppose we use the usual denominator $(n-1)$ in defining the sample variance and sample covariance for samples of size n:

$$\mathrm{Var}(x) = \frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \bar{x})^2$$

$$\mathrm{Cov}(x,y) = \frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Of course the correlation coefficient r is related to this covariance by

$$r = \frac{1}{s_x s_y}\left(\mathrm{Cov}(x,y)\right).$$

Then since $(a+b)^2 = a^2 + 2ab + b^2$, it follows that

$$\mathrm{Var(a + b)} = \mathrm{Var}(a) + \mathrm{Var}(b) + 2\mathrm{Cov}(a,b).$$

If we apply this to the usual simple linear regression setup, we obtain:

Proposition: The sample variance of the residuals d_i in a simple linear regression satisfies

$$\mathrm{Var}(d_i) = (1-r^2)\mathrm{Var}(y_i)$$

where $\mathrm{Var}(y_i)$ is the sample variance of the original response variable.

Proof: The line of regression may be written as

$$\hat{y} - \bar{y} = b_1(x - \bar{x})$$

where $b_1 = \frac{rs_y}{s_x}$. The residual of a point (x_i, y_i) is $d_i = y_i - \hat{y}_i$, so:

$$\begin{aligned}
\mathrm{Var}(d_i) &= \mathrm{Var}(y_i - \hat{y}_i) \\
&= \mathrm{Var}(y_i - (\bar{y} + b_1(x_i - \bar{x}))) \\
&= \mathrm{Var}((y_i - \bar{y}) - b_1(x_i - \bar{x})) \\
&= s_y^2 + b_1^2 s_x^2 - 2b_1\mathrm{Cov}(y_i - \bar{y}, x_i - \bar{x}) \\
&= s_y^2 + r^2\frac{s_y^2}{s_x^2}s_x^2 - 2r\frac{s_y}{s_x}(r s_x s_y) \\
&= s_y^2 - r^2 s_y^2 \\
&= (1-r^2)s_y^2.
\end{aligned}$$