Stat 410 Properties of a Regression Line

Dr. D. Scott

September 1, 2005

$$\widehat{Y} = b_0 + b_1 X \qquad \text{(regression prediction)}$$

but if no data collected around $X \approx 0 \ldots b_0$?
Re-centering

$$\widehat{Y} = b_0 + b_1 X \pm b_1 \bar{X}$$
$$\widehat{Y} = b_0 + b_1 \bar{X} + b_1 (X - \bar{X})$$

but $b_0 = \bar{Y} - b_1 \bar{X}$, so that

$$\widehat{Y} = \bar{Y} + b_1 (X - \bar{X}).$$

Notes: The point $(\bar{X}, \bar{Y})$ is on the regression line. If X is 1 unit more than $\bar{X}$, then $\widehat{Y}$ is b_1 units more than $\bar{Y}$.

Here are the maximum likelihood estimates of the variance, covariance, and correlation

$$var(x_i) = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2$$

$$cov = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}$$

$$cor(x_i, y_i) = \frac{cov(x_i, y_i)}{\sqrt{var(x_i)}\sqrt{var(y_i)}}$$

The correlation coefficient, ρ, is dimensionless and satisfies $-1 \leq \rho \leq 1$.

Properties of the residuals and predictions:

$$
\begin{aligned}
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} (Y_i - \widehat{Y}_i) \\
&= \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) \\
&= n\bar{Y} - nb_0 - nb_1\bar{X} \\
&= n\bar{Y} - n\left(\bar{Y} - b_1\bar{X}\right) - nb_1\bar{X} \\
&= 0\,.
\end{aligned}
$$

Hence,

$$\sum \widehat{Y}_i = \sum Y_i \qquad \text{(same average).}$$

Less obvious: e_i and X_i are uncorrelated.

$$
\begin{aligned}
\frac{1}{n}\sum(e_i - \bar{e})(X_i - \bar{X}) & \\
&= \frac{1}{n}\sum e_i X_i - \bar{e}\bar{X} \\
&= \frac{1}{n}\sum e_i X_i \quad \text{(since } \bar{e} = 0\text{)}
\end{aligned}
$$

continuing

$$
\begin{aligned}
&\frac{1}{n}\sum(Y_i - b_0 - b_1 X_i)X_i \\
&\qquad = \frac{1}{n}\sum X_i Y_i - b_0\bar{X} - \frac{1}{n}b_1\sum X_i^2 \\
&\qquad = \frac{1}{n}\sum X_i Y_i - (\bar{Y} - b_1\bar{X})\bar{X} - \frac{1}{n}b_1\sum X_i^2 \\
&\qquad = \frac{1}{n}\sum X_i Y_i - \bar{Y}\bar{X} + b_1\bar{X}^2 - \frac{1}{n}b_1\sum X_i^2 \\
&\qquad = cov(x_i, y_i) - b_1\left(\frac{1}{n}\sum X_i^2 - \bar{X}^2\right) \\
&\qquad = cov(x_i, y_i) - b_1\, var(x_i) \\
&\qquad = 0
\end{aligned}
$$

since $b_1 = cov(x_i, y_i)/var(x_i)$!

What is the big deal? If the two quantities X_i and Y_i are uncorrelated, then their covariance is also 0, and hence, so is b_1. Thus the best linear predictor is

$$\widehat{Y} = \bar{Y} + b_1(X - \bar{X}) = \bar{Y}.$$

Finally, e_i and $\widehat{Y}_i$ are uncorrelated.

$$\begin{aligned}\frac{1}{n}\sum e_i\widehat{Y}_i - \bar{e}\bar{\widehat{Y}} &= \frac{1}{n}\sum e_i\widehat{Y}_i = \frac{1}{n}\sum(e_i b_0 + b_1 e_i X_i) \\ &= \bar{e}b_0 + b_1\frac{1}{n}\sum e_i X_i = 0\,,\end{aligned}$$

since $\bar{e} = 0$ and e_i, x_i uncorrelated.