Stat 410 Properties of a Regression Line

Dr. D. Scott

September 1, 2005
\[ \hat{Y} = b_0 + b_1 X \] (regression prediction)
but if no data collected around \( X \approx 0 \)...
Re-centering
\[ \hat{Y} = b_0 + b_1 X \pm b_1 \bar{X} \]
\[ \hat{Y} = b_0 + b_1 \bar{X} + b_1 (X - \bar{X}) \]
but \( b_0 = \bar{Y} - b_1 \bar{X} \), so that
\[ \hat{Y} = \bar{Y} + b_1 (X - \bar{X}). \]

Notes: The point \((\bar{X}, \bar{Y})\) is on the regression line. If \( X \) is 1 unit more than \( \bar{X} \), then \( \hat{Y} \) is \( b_1 \) units more than \( \bar{Y} \).
Here are the maximum likelihood estimates of the variance, covariance, and correlation:

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

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

\[
\text{cor}(x_i, y_i) = \frac{\text{cov}(x_i, y_i)}{\sqrt{\text{var}(x_i)} \sqrt{\text{var}(y_i)}}
\]

The correlation coefficient, \( \rho \), is dimensionless and satisfies \(-1 \leq \rho \leq 1\).
Properties of the residuals and predictions:

\[
\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)
= \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)
= n\bar{Y} - nb_0 - nb_1 \bar{X}
= n\bar{Y} - n(\bar{Y} - b_1 \bar{X}) - nb_1 \bar{X}
= 0.
\]

Hence,

\[
\sum \hat{Y}_i = \sum Y_i \quad \text{(same average)}.
\]
Less obvious: $e_i$ and $X_i$ are uncorrelated.

\[
\frac{1}{n} \sum (e_i - \bar{e})(X_i - \bar{X}) = \frac{1}{n} \sum e_i X_i - \bar{e} \bar{X}
= \frac{1}{n} \sum e_i X_i \quad \text{(since } \bar{e} = 0)\]
continuing

\[
\frac{1}{n} \sum (Y_i - b_0 - b_1 X_i) X_i
\]

\[
= \frac{1}{n} \sum X_i Y_i - b_0 \bar{X} - \frac{1}{n} b_1 \sum X_i^2
\]

\[
= \frac{1}{n} \sum X_i Y_i - (\bar{Y} - b_1 \bar{X}) \bar{X} - \frac{1}{n} b_1 \sum X_i^2
\]

\[
= \frac{1}{n} \sum X_i Y_i - \bar{Y} \bar{X} + b_1 \bar{X}^2 - \frac{1}{n} b_1 \sum X_i^2
\]

\[
= \text{cov}(x_i, y_i) - b_1 \left( \frac{1}{n} \sum X_i^2 - \bar{X}^2 \right)
\]

\[
= \text{cov}(x_i, y_i) - b_1 \text{var}(x_i)
\]

\[
= 0
\]

since \( b_1 = \text{cov}(x_i, y_i)/\text{var}(x_i) \)!
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

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

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

$$\frac{1}{n} \sum e_i \hat{Y}_i - \bar{e} \bar{\hat{Y}} = \frac{1}{n} \sum e_i \hat{Y}_i = \frac{1}{n} \sum (e_i b_0 + b_1 e_i X_i)$$

$$= \bar{e} b_0 + b_1 \frac{1}{n} \sum e_i X_i = 0,$$

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