A simple least-squares model

Let \((x_i, y_i), \ i = 1, \ldots, n\) be a given set of points in the plane. We will use the notation:

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \quad s_{xx} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{y}) x_i = \bar{x}^2 - \bar{y}^2, \\
\]

\[
s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{y}) y_i = \frac{1}{n} \sum_{i=1}^{n} x_i (y_i - \bar{y}) = \bar{x} \bar{y} - \bar{y} y. \\
\]

For a fixed value of \(\beta_1\), the mean squared vertical deviation from the line \(y = \beta_1 x + \beta_0\),

\[
\frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2,
\]

is a minimum when \(\beta_0 = \bar{y} - \beta_1 \bar{x}\). The minimum is \(\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y} - \beta_1 (x_i - \bar{x}))^2 = s_{yy} - 2 s_{xy} \beta_1 + s_{xx} \beta_1^2\).

In turn, this minimum is a minimum when \(\beta_1 = \frac{s_{xy}}{s_{xx}}\). The minimum of the minimum is \(\frac{s_{xx}s_{yy} - s_{xy}^2}{s_{xx}}\).

Suppose now that \(y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, \ldots, n\), where \(\epsilon_i\) are iid random variables with mean 0 and variance \(\sigma^2\) (noise). Then each \(y_i\), \(\beta_1 = s_{xy}/s_{xx}\), and \(\beta_0 = \bar{y} - \beta_1 \bar{x}\) are random variables as well, and their dependence on \(\epsilon_i\) is linear. Note that \(y_i\) are independent, but \(\beta_1\) and \(\beta_0\) are not.

**Claim.** The least-squares coefficient estimators are unbiased, \(\mathbb{E}[\hat{\beta}_0] = \beta_0\) and \(\mathbb{E}[\hat{\beta}_1] = \beta_1\).

**Proof:**

\[
\mathbb{E}[y_i] = \mathbb{E}[\beta_0 + \beta_1 x_i + \epsilon_i] = \beta_0 + \beta_1 x_i + 0 \\
\mathbb{E}[\bar{y}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[y_i] = \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \bar{x} \\
\mathbb{E}[s_{xy}] = \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^{n} (x_i - \bar{x}) y_i \right] = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{y}) (\beta_0 + \beta_1 x_i) = \beta_1 \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) x_i = \beta_1 s_{xx} \\
\mathbb{E}[\hat{\beta}_1] = \frac{1}{s_{xx}} \mathbb{E}[s_{xy}] = \frac{1}{s_{xx}} \beta_1 s_{xx} = \beta_1 \\
\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{y} - \beta_1 \bar{x}] = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0 \\
\]

**Claim.** The variances are \(\text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} \frac{\bar{x}^2}{s_{xx}}\) and \(\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n} \frac{1}{s_{xx}}\).

**Proof:**

\[
\text{Var}(y_i) = \text{Var}(\epsilon_i) = \sigma^2 \\
\text{Var}(s_{xy}) = \frac{1}{n^2} \text{Var} \left( \sum_{i=1}^{n} (x_i - \bar{x}) y_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} (x_i - \bar{y})^2 \sigma^2 = \frac{\sigma^2}{n} s_{xx} \\
\text{Var}(\hat{\beta}_1) = \frac{1}{s_{xx}^2} \text{Var}(s_{xy}) = \frac{\sigma^2}{n} \frac{1}{s_{xx}} \\
\text{Var}(\hat{\beta}_0) = \text{Var} \left( \sum_{i=1}^{n} \left[ 1 - (x_i - \bar{x}) \frac{\bar{y}}{s_{xx}} \right] y_i \right) = \frac{\sigma^2}{n^2} \sum_{i=1}^{n} \left[ 1 - (x_i - \bar{x}) \frac{\bar{y}}{s_{xx}} \right]^2 \frac{\bar{y}^2}{s_{xx}^2} = \frac{\sigma^2}{n} \frac{\bar{x}^2}{s_{xx}} \\
\]

Since \(\hat{\beta}_0 + \bar{x} \hat{\beta}_1 = \bar{y}\), we have \(\text{Var}(\hat{\beta}_0) + \bar{x}^2 \text{Var}(\hat{\beta}_1) + 2 \bar{x} \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \sigma^2 \frac{\bar{x}^2}{n} + 2 \bar{x} \frac{\sigma^2}{n} \frac{\bar{x}^2}{s_{xx}}\) and so \(\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2}{n} \frac{\bar{x}}{s_{xx}}\).