Proof 9-2
Larson Hogstrom - Math E156 - 2014

Least-squares regression

You have values x_i of a "predictor" and matching values y_i of a "response." Your goal is to minimize the sum of squares of the prediction errors,

$$g(a,b) = \sum_{i=1}^{n} (a + bx_i - y_i)^2.$$

Prove that

$$b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, a = \overline{y} - b\overline{x}.$$

Solution:

$$\frac{dg}{da} = 2\sum_{i=1}^{n} (a + bx_i - y_i)$$

$$0 = \sum_{i=1}^{n} (a + bx_i - y_i)$$

$$-na = b\sum_{i=1}^{n} (x_i) + \sum_{i=1}^{n} (y_i)$$

$$-na = bn\overline{x} + n\overline{y}$$

$$a = \overline{y} - b\overline{x}$$

$$\frac{dg}{db} = 2\sum_{i=1}^{n} x_i(a + bx_i - y_i)$$

$$\frac{dg}{db} = 2\sum_{i=1}^{n} x_i(\overline{y} - b\overline{x} + bx_i - y_i)$$

Then add: $-\sum_{i=1}^{n} \overline{x}(a + bx_i + y_i)$

$$0 = \sum_{i=1}^{n}(x_i - \overline{x})(\overline{y} - b\overline{x} + bx_i - y_i)$$

$$0 = \sum_{i=1}^{n}(x_i - \overline{x})(bx_i - b\overline{x}) - \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

$$0 = b\sum_{i=1}^{n}(x_i - \overline{x})^2 - \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$