Least squares

Four preliminary facts:

1. If \( u = (u_1, \ldots, u_n) \) is a vector of real numbers then \( \sum_{i=1}^{n}(u_i - \bar{u}) = 0 \).

2. The function \( h(x) = \sum_{i=1}^{n}(u_i - x)^2 \) is minimized at \( x = \bar{u} \).

3. The function \( f(x) = ax^2 - 2bx + c \) is minimized at \( x = b/a \) and \( f(b/a) = c - b^2/a \) when \( a > 0 \).

4. \( \sum_{i=1}^{n}(u_i - \bar{u})^2 = \sum_{i=1}^{n}u_i^2 - n\bar{u}^2 \)

Let \((x_1, y_1), \ldots, (x_n, y_n)\) be \( n \) fixed points. The least squares line \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \) is the solution to the following problem: Minimize over \( \beta_0 \) and \( \beta_1 \)

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

We will now show that the solution is given by

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}
\]

Now for a fixed \( \beta_1 \) equation 1 which is minimized by \( \beta_0 = \bar{y} - \beta_1 \bar{x} \) by preliminary fact 2. This also proves equation 2. So to solve equation 1 it is enough to minimize over \( \beta_1 \)

\[
\sum_{i=1}^{n}(y_i - \bar{y} - \beta_1 (x_i - \bar{x}))^2
\]

But note

\[
\sum_{i=1}^{n}(y_i - \bar{y} - \beta_1 (x_i - \bar{x}))^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})
\]

\[
= a\beta_1^2 - 2b\beta_1 + c
\]

where

\[
a = \sum_{i=1}^{n}(x_i - \bar{x})^2, \ b = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \text{ and } c = \sum_{i=1}^{n}(y_i - \bar{y})^2
\]

By preliminary fact 3 this quadratic in \( \beta_1 \) is minimize at \( b/a \) so equation 3 follows.

Using equation 2 we can write the least squares line as

\[
\hat{y} = \bar{y} + \hat{\beta}_1 (x - \bar{x})
\]

This has two important consequences. First the point \( (\bar{x}, \bar{y}) \) must lie on the least squares line. Secondly if \( y_i - \hat{y}_i = y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}) \) is the \( i \)th residual then we have by preliminary fact 1 that

\[
\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})
\]

\[
= 0 + 0
\]
Finally by the second part of preliminary fact 3 we have that

\[ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = c - b^2/a \]

\[ = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \left( \frac{\sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2 \]

\[ = \sum_{i=1}^{n} (y_i - \bar{y})^2 \left( 1 - \frac{\left( \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}) \right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2} \right) \]

\[ = \sum_{i=1}^{n} (y_i - \bar{y})^2 (1 - \hat{\rho}_{x,y}^2) \]

where

\[ \hat{\rho}_{x,y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{S_{x,y}}{\sqrt{S_{x,x}S_{y,y}}} \] (5)

is the sample correlation coefficient.

From now on \( \sum_{i=1}^{n} \) will just be written as \( \sum \). Note the equation on the top of the page can be rewritten as

\[ \sum (y_i - \bar{y})^2 = \hat{\rho}_{x,y}^2 \sum (y_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \] (6)

An equivalent form of this equation is

\[ \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \] (7)

To see this note that

\[ \sum (\hat{y}_i - \bar{y})^2 = \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}))^2 \]

\[ = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 \]

\[ = \frac{S_{x,y}^2}{S_{x,x}} \]

\[ = \frac{S_{x,y}^2 S_{y,y}}{S_{x,x} S_{y,y}} \]

\[ = \hat{\rho}_{x,y}^2 \sum (y_i - \bar{y})^2 \]

Using preliminary fact 4 equation 7 can be rewritten as

\[ \sum y_i^2 = ny^2 + \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \] (8)

or using notation given in class this can be written as

\[ TSS = SSR(\beta_0) + SSR(\beta_1|\beta_0) + RSS \] (9)

The analogous version for equation 7 is

\[ TCSS = SSR(\beta_1|\beta_0) + RSS \] (10)