LEAST SQUARES REGRESSION

Assumption: (Simple Linear Model, Version 1)

1. $E(Y|x) = \eta_0 + \eta_1 x$ (linear mean function)

Goal: To estimate $\eta_0$ and $\eta_1$ (and later $\sigma^2$) from data.

Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Notation:

- The estimates of $\eta_0$ and $\eta_1$ will be denoted by $\hat{\eta}_0$ and $\hat{\eta}_1$, respectively. They are called the ordinary least squares (OLS) estimates of $\eta_0$ and $\eta_1$.

- $\hat{E}(Y|x) = \hat{\eta}_0 + \hat{\eta}_1 x = \hat{y}$

- The line $y = \hat{\eta}_0 + \hat{\eta}_1 x$ is called the ordinary least squares (OLS) line.

- $\hat{y}_i = \hat{\eta}_0 + \hat{\eta}_1 x_i$ (i\textsuperscript{th} fitted value or i\textsuperscript{th} fit)

- $\hat{e}_i = y_i - \hat{y}_i$ (i\textsuperscript{th} residual)

Set-up:

- Consider lines $y = h_0 + h_1 x$.

- $d_i = y_i - (h_0 + h_1 x_i)$

- $\hat{h}_0$ and $\hat{h}_1$ will be the values of $h_0$ and $h_1$ that minimize $\sum d_i^2$. 
More Notation:

- \( \text{RSS}(h_0, h_1) = \sum d_i^2 \) (for Residual Sum of Squares).

- \( \text{RSS} = \text{RSS}(\hat{h}_0, \hat{h}_1) = \sum \hat{e}_i^2 \) -- "the" Residual Sum of Squares (i.e., the minimal residual sum of squares)

Solving for \( \hat{h}_0 \) and \( \hat{h}_1 \):

- We want to minimize the function \( \text{RSS}(h_0, h_1) = \sum d_i^2 = \sum [y_i - (h_0 + h_1x_i)]^2 \)

- Visually, there is no maximum. [See Demos]

- \( \text{RSS}(h_0, h_1) \geq 0 \)

- Therefore if there is a critical point, minimum occurs there.

To find critical points:

\[
\begin{align*}
\frac{\partial \text{RSS}}{\partial h_0}(h_0, h_1) &= \sum 2[y_i - (h_0 + h_1x_i)](-1) \\
\frac{\partial \text{RSS}}{\partial h_1}(h_0, h_1) &= \sum 2[y_i - (h_0 + h_1x_i)](-x_i)
\end{align*}
\]

So \( \hat{h}_0, \hat{h}_1 \) must satisfy the *normal equations*

\[
\begin{align*}
\text{(i)} & \quad \frac{\partial \text{RSS}}{\partial h_0}(\hat{h}_0, \hat{h}_1) = \sum (-2)[y_i - (\hat{h}_0 + \hat{h}_1x_i)] = 0 \\
\text{(ii)} & \quad \frac{\partial \text{RSS}}{\partial h_1}(\hat{h}_0, \hat{h}_1) = \sum (-2)[y_i - (\hat{h}_0 + \hat{h}_1x_i)]x_i = 0
\end{align*}
\]

Cancelling the -2's and recalling that \( \hat{e}_i = y_i - \hat{y}_i \):

\[
\begin{align*}
\text{(i)}' & \quad \sum \hat{e}_i = 0 \\
\text{(ii)}' & \quad \sum \hat{e}_i x_i = 0
\end{align*}
\]

In words:

\[
\begin{align*}
\text{(i)}' & \quad \sum y_i = \sum \hat{y}_i \\
\text{(ii)}' & \quad \sum x_i y_i = \sum x_i \hat{y}_i
\end{align*}
\]
Visually:

Note that (i) implies $\bar{e}_i = 0$

(since sample mean of the $\hat{e}_i$'s is zero)

To solve the normal equations:

(i) $\Rightarrow \sum y_i - \hat{\beta}_h \sum x_i = 0$

$\Rightarrow n\bar{y} - n\hat{\beta}_h (n\bar{x}) = 0$

$\Rightarrow \bar{y} - \hat{\beta}_h \bar{x} = 0$

Consequences:

- Can solve for $\hat{\beta}_h$ once we find $\hat{\beta}_h = \bar{y} - \hat{\beta}_h \bar{x}$

- $\bar{y} = \hat{\beta}_h \bar{x}$, which says:

Note analogies to bivariate normal mean line:

- $\alpha_{Y|x} = E(Y|X) - \beta_{Y|x} E(X)$  (equation 4.14)

- $(\mu_{X,Y})$ lies on the (population) mean line (Problem 4.7)

(ii) $\Rightarrow$ (substituting $\hat{\beta}_h = \bar{y} - \hat{\beta}_h \bar{x}$)

$\Rightarrow \sum [y_i - (\bar{y} - \hat{\beta}_h \bar{x} + \hat{\beta}_h x_i)]x_i = 0$

$\Rightarrow \sum [(y_i - \bar{y}) - \hat{\beta}_h (x_i - \bar{x})]x_i = 0$

$\Rightarrow \sum x_i (y_i - \bar{y}) - \hat{\beta}_h \sum x_i (x_i - \bar{x}) = 0$

$\Rightarrow \hat{\beta}_h = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})}$

Notation:

$SXX = \sum x_i (x_i - \bar{x})$  $SYY = \sum y_i (y_i - \bar{y})$

$SXY = \sum x_i (y_i - \bar{y})$

So for short: $\hat{\beta}_h = \frac{SXY}{SXX}$
Useful identities:

• $SXX = \sum (x_i - \bar{x})^2$
• $SXY = \sum (x_i - \bar{x})(y_i - \bar{y})$
• $SXY = \sum (x_i - \bar{x})y_i$
• $SYY = \sum (y_i - \bar{y})^2$

Proof of (1):

$$\sum (x_i - \bar{x})^2 = \sum [x_i (x_i - \bar{x}) - \bar{x}(x_i - \bar{x})] = \sum x_i (x_i - \bar{x}) - \bar{x}\sum (x_i - \bar{x}),$$

and

$$\sum (x_i - \bar{x}) = \sum x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

(Try the others yourself!)

Summarize:

$$\hat{\beta}_1 = \frac{SXY}{SXX}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \frac{SXY}{SXX} \bar{x}$$

Connection with Sample Correlation Coefficient

Recall: The sample correlation coefficient

$$r = r(x,y) = \hat{\rho}(x,y) = \frac{\text{cov}(x,y)}{sd(x)sd(y)}$$

(Note: everything here calculated from sample.)

Note that:

$$\text{cov}(x,y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} SXY$$

$$[sd(x)]^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{1}{n-1} SXX$$

$$[sd(y)]^2 = \frac{1}{n-1} SYY \quad (similarly)$$

Therefore:
More on r: Recall:

\[ \tilde{y}_i = \hat{\eta}_0 + \hat{\eta}_1 x_i \]

**Fits**

\[ \hat{\eta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

**Residuals**

\[ \hat{\epsilon}_i = y_i - \hat{\tilde{y}}_i = y_i - (\hat{\eta}_0 + \hat{\eta}_1 x_i) \]

\[ \text{RSS} = \text{RSS}(\hat{\eta}_0, \hat{\eta}_1) = \sum \hat{\epsilon}_i^2 \text{ -- "the" Residual Sum of Squares (i.e., the minimal residual sum of squares)} \]

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

i.e., the point \((\bar{x}, \bar{y})\) is on the least squares line.

Recall and note analogy: For a bivariate normal distribution,

\[ \text{E}(Y|X = x) = \alpha_{yx} + \beta_{yx} x \quad \text{(equation 4.13)} \]

where \[ \beta_{yx} = \rho \frac{\sigma_y}{\sigma_x} \]
Calculate:

\[ \text{RSS} = \sum \hat{e}_i^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \]

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

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

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

\[ = \text{SYY} - 2 \frac{\text{SXY}}{\text{SXX}} \text{SXY} + \left( \frac{\text{SXY}}{\text{SXX}} \right)^2 \text{SXX} \]

\[ = \text{SYY} - \frac{(\text{SXY})^2}{\text{SXX}} \]

\[ = \text{SYY} \left[ 1 - \frac{(\text{SXY})^2}{(\text{SXX})(\text{SYY})} \right] \]

\[ = \text{SYY}(1 - r^2) \]

Thus

\[ 1 - r^2 = \frac{\text{RSS}}{\text{SYY}}, \]

so

\[ r^2 = 1 - \frac{\text{RSS}}{\text{SYY}} = \frac{\text{SYY} - \text{RSS}}{\text{SYY}} \]

Interpretation:

\[ \text{SYY} = \sum (y_i - \bar{y})^2 \]

is a measure of the total variability of the \( y_i \)'s from \( \bar{y} \).

\[ \text{RSS} = \sum \hat{e}_i^2 \]

is a measure of the variability in \( y \) remaining after conditioning on \( x \) (i.e., after regressing on \( x \)).

So

\[ \text{SYY} - \text{RSS} \]

is a measure of the amount of variability of \( y \) accounted for by conditioning (i.e., regressing) on \( x \).

Thus

\[ r^2 = \frac{\text{SYY} - \text{RSS}}{\text{SYY}} \]

is the proportion of the total variability in \( y \) accounted for by regressing on \( x \).
Note: One can show (details left to the interested student) that $SYY - RSS = \sum (\hat{y}_i - \bar{y})^2$ and $\bar{y}_i = \bar{y}$, so that in fact $r^2 = \frac{\text{var}(\hat{y})}{\text{var}(y)}$, the proportion of the sample variance of $y$ accounted for by regression on $x$. 