LEAST SQUARES REGRESSION

Assumptions for the Simple Linear Model:
1. \( E(Y|x) = \eta_0 + \eta_1 x \) (linear mean function)
2. \( \text{Var}(Y|x) = \sigma^2 \) (constant variance)

Equivalent form of (2):
2': \( \text{Var}(e|x) = \sigma^2 \) (constant error variance)

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.
- \( \hat{y}_i = \hat{\eta}_0 + \hat{\eta}_1 x_i \) (\( i^{th} \) fitted value or \( i^{th} \) fit)
- \( \hat{e}_i = y_i - \hat{y}_i \) (\( i^{th} \) residual)

Set-up:
1. Consider lines \( y = h_0 + h_1 x. \)
2. \( d_i = y_i - (h_0 + h_1 x_i) \)
3. \( \hat{h}_0 \) 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 \( \text{RSS}(h_0, h_1) = \sum d_i^2 = \sum [y_i - (h_0 + h_1 x_i)]^2 \)
- [Recall Demo]

- Visually, there is no maximum.
- \( \text{RSS}(h_0, h_1) \geq 0 \)
- Therefore if there is a critical point, minimum occurs there.

To find critical points:
\[
\frac{\partial \text{RSS}}{\partial h_0} (h_0, h_1) = \sum 2[y_i - (h_0 + h_1 x_i)](-1)
\]
\[
\frac{\partial \text{RSS}}{\partial h_1} (h_0, h_1) = \sum 2[y_i - (h_0 + h_1 x_i)](-x_i)
\]

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

(i) \( \frac{\partial \text{RSS}}{\partial h_0} (\hat{h}_0, \hat{h}_1) = \sum (-2)[y_i - (\hat{h}_0 + \hat{h}_1 x_i)] = 0 \)

(ii) \( \frac{\partial \text{RSS}}{\partial h_1} (\hat{h}_0, \hat{h}_1) = \sum (-2)[y_i - (\hat{h}_0 + \hat{h}_1 x_i)]x_i = 0 \)

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

(i)' \( \sum \hat{e}_i = 0 \)

(ii)' \( \sum \hat{e}_i x_i = 0 \)

In words:

Visually:

Note: (i)' implies \( \bar{e} = 0 \) (sample mean of the \( \hat{e}_i \)'s is zero)

To solve the normal equations:

(i) \( \Rightarrow \sum y_i - \sum \hat{h}_0 - \hat{h} \sum x_i \)

\[ \Rightarrow n \bar{y} - n \hat{h}_0 - \hat{h}(n \bar{x}) = 0 \]

\[ \Rightarrow \bar{y} - \hat{h}_0 - \hat{h} \bar{x} = 0 \]

Consequences:

\( \bullet \) Can use to solve for \( \hat{h}_0 \) once we find \( \hat{h}_1 \): \( \hat{h}_0 = \bar{y} - \hat{h}_1 \bar{x} \)

\( \bullet \) \( \bar{y} = \hat{h}_0 + \hat{h}_1 \bar{x} \), which says:

Note analogies to bivariate normal mean line:

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

\( \bullet \) \( (\mu_X, \mu_Y) \) lies on the (population) mean line (Problem 4.7)

(ii') \( \Rightarrow \) (substituting \( \hat{h}_0 = \bar{y} - \hat{h}_1 \bar{x} \))

\[ \sum [y_i - (\bar{y} - \hat{h}_1 \bar{x} + \hat{h}_1 x_i)]x_i = 0 \]

\[ \Rightarrow \sum [(y_i - \bar{y}) - \hat{h}_1 (x_i - \bar{x})]x_i = 0 \]

\[ \Rightarrow \sum x_i (y_i - \bar{y}) - \hat{h}_1 \sum x_i (x_i - \bar{x}) = 0 \]
\[ \Rightarrow \hat{\eta}_h = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})} \]

Notation:
- \( SXX = \sum x_i (x_i - \bar{x}) \)
- \( SXY = \sum x_i (y_i - \bar{y}) \)
- \( SYY = \sum y_i (y_i - \bar{y}) \)

So for short:
\[ \hat{\eta}_h = \frac{SXY}{SXX} \]

Useful identities:
1. \( SXX = \sum (x_i - \bar{x})^2 \)
2. \( SXY = \sum (x_i - \bar{x})(y_i - \bar{y}) \)
3. \( SXY = \sum (x_i - \bar{x})y_i \)
4. \( SYY = \sum (y_i - \bar{y})^2 \)

Proof of (1):
\[
\sum (x_i - \bar{x})^2 = \sum [x_i (x_i - \bar{x}) - \bar{x} (x_i - \bar{x})] \\
= \sum x_i (x_i - \bar{x}) - \bar{x} \sum (x_i - \bar{x}) \\
\text{and} \\
\sum (x_i - \bar{x}) = \sum x_i - n \bar{x} \\
= n \bar{x} - n \bar{x} = 0
\]
(Try proving (2) - (4) yourself!)

Summarize:
\[ \hat{\eta}_h = \frac{SXY}{SXX} \]
\[
\hat{\eta}_0 = \bar{y} - \hat{\eta}_h \bar{x} \\
= \bar{y} - \frac{SXY}{SXX} \bar{x}
\]

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 that everything here is calculated from the sample.)
Note that:
\[
cov(x, y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} S_{XY}
\]
\[
[\text{sd}(x)]^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{1}{n-1} S_{XX}
\]
and similarly,
\[
[\text{sd}(y)]^2 = \frac{1}{n-1} S_{YY}
\]
Therefore:
\[
 r^2 = \frac{[\text{cov}(x, y)]^2}{[\text{sd}(x)]^2[\text{sd}(y)]^2} = \frac{\left( \frac{1}{n-1} \right)^2 (S_{XY})^2}{\left( \frac{1}{n-1} S_{XX} \right) \left( \frac{1}{n-1} S_{YY} \right)} = \frac{(S_{XY})^2}{(S_{XX})(S_{YY})}
\]
Also,
\[
 r \frac{\text{sd}(y)}{\text{sd}(x)} = \frac{\text{cov}(x, y)}{\text{sd}(x)\text{sd}(y)} = \frac{\text{cov}(x, y)}{\text{sd}(x)^2} = \frac{1}{n-1} S_{XY} = \frac{1}{n-1} S_{XX}
\]
\[
\frac{S_{XY}}{S_{XX}} = \hat{\eta}
\]
For short:
\[
\hat{\eta} = r \frac{s_y}{s_x}
\]
Recall and note the analogy: For a bivariate normal distribution,
\[ E(Y|X = x) = \alpha_{yx} + \beta_{yx}x \]  
\[(\text{equation 4.13)}\]

where \( \beta_{yx} = \frac{\sigma_y}{\sigma_x} \)

More on \( r \):

Recall:

- Fits: \( \hat{y}_i = \hat{\eta}_0 + \hat{\eta}_1 x_i \)
- Residuals: \( \hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\eta}_0 + \hat{\eta}_1 x_i) \)

\[ \text{RSS}(\hat{\eta}_0, \hat{\eta}_1) = \sum d_i^2 \]

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

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

Calculate:

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

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

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

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

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

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

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

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

Thus

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

so

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

- $SYY = \sum (y_i - \bar{y})^2$ is a measure of the total variability of the $y_i$'s from $\bar{y}$.
- $RSS = \sum \hat{e}_i^2$ is a measure of the variability in $y$ remaining *after* conditioning on $x$ (i.e., after regressing on $x$).

So

$SYY - RSS$ is a measure of the amount of variability of $y$ *accounted for* by conditioning (i.e., regressing) on $x$.

Thus

$$r^2 = \frac{SYY - RSS}{SYY}$$

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 $r^2 = \frac{\text{var}(\hat{y}_i)}{\text{var}(y_i)}$, the proportion of the sample variance of $y$

accounted for by regression on $x$. 