LEAST SQUARES REGRESSION

Our only assumption so far for the Simple Linear Model:

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

   [Picture]

Goal: To estimate η_0 and η_1 from data.

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

Notation:

- The estimates of η_0 and η_1 will be denoted by $\hat{\eta}_0$ and $\hat{\eta}_1$, respectively. They are called the *ordinary least squares (OLS) estimates* of η_0 and η_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. $\hat{\eta}_0$ and $\hat{\eta}_1$ will be the values of h_0 and h_1 that minimize $\sum d_i^2$.

More Notation:

- $RSS(h_0, h_1) = \sum d_i^2$ (for Residual Sum of Squares).
- $RSS = RSS(\hat{\eta}_0, \hat{\eta}_1) = \sum \hat{e}_i^{\,2}$ -- "the" Residual Sum of Squares (i.e., the minimal residual sum of squares)

Solving for $\hat{\eta}_0$ and $\hat{\eta}_1$:

- We want to minimize the function $RSS(h_0, h_1) = \sum d_i^2 = \sum [y_i - (h_0 + h_1 x_i)]^2$
- [Recall Demo]

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

To find critical points:

$$\frac{\partial RSS}{\partial h_0}(h_0,h_1) = \sum 2[y_i - (h_0 + h_1 x_i)](-1)$$

$$\frac{\partial RSS}{\partial h_1}(h_0,h_1) = \sum 2[y_i - (h_0 + h_1 x_i)](-x_i)$$

So $\hat{\eta}_0, \hat{\eta}_1$ must satisfy the *normal equations*

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

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

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 $\overline{\hat{e}_i} = 0$ (sample mean of the $\hat{e}_i$'s is zero)

To solve the normal equations:

(i) $\Rightarrow \sum y_i - \sum \hat{\eta}_0 - \hat{\eta}_1 \sum x_i$

$\Rightarrow n\bar{y} - n\hat{\eta}_0 - \hat{\eta}_1(n\bar{x}) = 0$

$\Rightarrow \bar{y} - \hat{\eta}_0 - \hat{\eta}_1 \bar{x} = 0$

Consequences:

- Can use to solve for $\hat{\eta}_0$ once we find $\hat{\eta}_1$: $\hat{\eta}_0 = \bar{y} - \hat{\eta}_1 \bar{x}$
- $\bar{y} = \hat{\eta}_0 + \hat{\eta}_1 \bar{x}$, which says:

Note analogies to bivariate normal mean line:

- $\alpha_{Y|x} = E(Y) - \beta_{Y|x}E(X)$ (equation 4.14)
- $(\mu_{X.}, \mu_Y)$ lies on the (population) mean line (Problem 4.7)

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

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

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

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

$$\Rightarrow \hat{\eta}_1 = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})}$$

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:

$$\hat{\eta}_1 = \frac{SXY}{SXX}$$

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):

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

and

$$\begin{aligned} \sum (x_i - \bar{x}) &= \sum x_i - n\bar{x} \\ &= n\bar{x} - n\bar{x} = 0 \end{aligned}$$

(Try proving (2) - (4) yourself!)

Summarize:

$$\hat{\eta}_1 = \frac{SXY}{SXX}$$

$$\begin{aligned} \hat{\eta}_0 &= \bar{y} - \hat{\eta}_1 \bar{x} \\ &= \bar{y} - \frac{SXY}{SXX} \bar{x} \end{aligned}$$

Connection with Sample Correlation Coefficient

Recall: The *sample correlation coefficient*

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

(Note that everything here is calculated from the sample.)

Note that:

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

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

$$= \frac{1}{n-1}\mathrm{SXX}$$

and similarly,

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

Therefore:

$$\mathrm{r}^2 = \frac{[\mathrm{c\hat{o}v}(x,y)]^2}{[sd(x)]^2[sd(y)]^2}$$

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

$$= \frac{(SXY)^2}{(SXX)(SYY)}$$

Also,

$$\mathrm{r}\frac{sd(y)}{sd(x)} = \frac{\mathrm{c\hat{o}v}(x,y)}{sd(x)sd(y)}\,\frac{sd(y)}{sd(x)}$$

$$= \frac{\mathrm{c\hat{o}v}(x,y)}{sd(x)^2}$$

$$= \frac{\frac{1}{\mathrm{n-1}}SXY}{\frac{1}{\mathrm{n-1}}SXX}$$

$$= \frac{SXY}{SXX} = \hat{\eta}_1$$

For short:

$$\hat{\eta}_1 = \mathrm{r}\frac{s_y}{s_x}$$

Recall and note the analogy: For a bivariate normal distribution,

$$\mathrm{E}(\mathrm{Y}|\mathrm{X} = \mathrm{x}) = \alpha_{\mathrm{Y|x}} + \beta_{\mathrm{Y|X}}\mathrm{x} \qquad \text{(equation 4.13)}$$

where $\beta_{\mathrm{Y|X}} = \rho\frac{\sigma_y}{\sigma_x}$

More on r:

Recall: [Picture]

Fits $\hat{y}_i = \hat{\eta}_0 + \hat{\eta}_1 x_i$

Residuals $\hat{e}_i = y_i - \hat{y}_i$

$= y_i - (\hat{\eta}_0 + \hat{\eta}_1 x_i)$

$\text{RSS}(h_0, h_1) = \sum d_i^2$

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

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

Calculate:

$$
\begin{aligned}
\text{RSS} = \sum \hat{e}_i^{\,2} &= \sum[y_i - (\hat{\eta}_0 + \hat{\eta}_1 x_i)]^2 \\
&= \sum[y_i - (\bar{y} - \hat{\eta}_1 \bar{x}) - \hat{\eta}_1 x_i]^2 \\
&= \sum[(y_i - \bar{y}) - \hat{\eta}_1 (x_i - \bar{x})]^2 \\
&= \sum[(y_i - \bar{y})^2 - 2\hat{\eta}_1 (x_i - \bar{x})(y_i - \bar{y}) + \hat{\eta}_1^{\,2}(x_i - \bar{x})^2] \\
&= \sum(y_i - \bar{y})^2 - 2\hat{\eta}_1 \sum (x_i - \bar{x})(y_i - \bar{y}) + \hat{\eta}_1^{\,2} \sum (x_i - \bar{x})^2 \\
&= \text{SYY} - 2\frac{SXY}{SXX}\text{SXY} + \left(\frac{SXY}{SXX}\right)^2 \text{SXX} \\
&= \text{SYY} - \frac{(SXY)^2}{SXX} \\
&= \text{SYY}\left[1 - \frac{(SXY)^2}{(SXX)(SYY)}\right] \\
&= \text{SYY}(1 - r^2)
\end{aligned}
$$

Thus

$$1 - r^2 = \frac{RSS}{SYY},$$

so

$$r^2 = 1 - \frac{RSS}{SYY} = \frac{SYY - RSS}{SYY}$$

Interpretation: [Picture]

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

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 $\overline{\hat{y}_i} = \bar{y}$, so that in fact $r^2 = \frac{\hat{\text{var}}(\hat{y}_i)}{\hat{\text{var}}(y_i)}$, the proportion of the sample variance of y accounted for by regression on x.