

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):

$$2': \text{Var}(e|x) = \sigma^2 \quad (\text{constant error variance})$$

[Picture]

Goal: To estimate η_0 and η_1 (and later σ^2) from data.

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

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:

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

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

- 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:

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

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

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

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

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

In words:

Visually:

Note: (i)' implies $\bar{\hat{e}} = 0$ (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)
- (μ_X, μ_Y) lies on the 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:

- $SXX = \sum x_i(x_i - \bar{x})$
- $SXY = \sum x_i(y_i - \bar{y})$
- $SY Y = \sum y_i(y_i - \bar{y})$

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. $SY Y = \sum (y_i - \bar{y})^2$

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*

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

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

Note that:

$$\begin{aligned}\hat{\text{cov}}(x,y) &= \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n-1} SXY \\ [\text{sd}(x)]^2 &= \frac{1}{n-1} \sum (x_i - \bar{x})^2 \\ &= \frac{1}{n-1} SXX\end{aligned}$$

and similarly,

$$[\text{sd}(y)]^2 = \frac{1}{n-1} SY Y$$

Therefore:

$$\begin{aligned}r^2 &= \frac{[\hat{\text{cov}}(x,y)]^2}{[\text{sd}(x)]^2[\text{sd}(y)]^2} \\ &= \frac{\left(\frac{1}{n-1}\right)^2 (SXY)^2}{\left(\frac{1}{n-1} SXX\right)\left(\frac{1}{n-1} SY Y\right)} \\ &= \frac{(SXY)^2}{(SXX)(SY Y)}\end{aligned}$$

Also,

$$\begin{aligned}r \frac{\text{sd}(y)}{\text{sd}(x)} &= \frac{\hat{\text{cov}}(x,y)}{\text{sd}(x)\text{sd}(y)} \frac{\text{sd}(y)}{\text{sd}(x)} \\ &= \frac{\hat{\text{cov}}(x,y)}{\text{sd}(x)^2} \\ &= \frac{\frac{1}{n-1} SXY}{\frac{1}{n-1} SXX} \\ &= \frac{SXY}{SXX} = \hat{\eta}_h\end{aligned}$$

For short:

$$\hat{\eta}_h = r \frac{s_y}{s_x}$$

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

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

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

More on r:

Recall: [Picture}

$$\begin{aligned} \text{Fits} \quad \hat{y}_i &= \hat{\eta}_0 + \hat{\eta}_1 x_i \\ \text{Residuals} \quad \hat{e}_i &= y_i - \hat{y}_i \\ &= y_i - (\hat{\eta}_0 + \hat{\eta}_1 x_i) \end{aligned}$$

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

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

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

Calculate:

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

Thus

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

so

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

Interpretation:

[Picture]

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