Simple Linear Regression (SLR)

In this situation we have a response variable (y) and one explanatory variable (x). In this, the model reduces to:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

In this the data consists of n paired observations (x_i, y_i) for $i = 1, 2, \ldots, n$. This is best summarized by a scatterplot (plot of y vs. x). Such a plot provides information about the direction and strength of the relationship between x and y. This relationship can also be formally quantified with the correlation coefficient:

$$\rho = Corr(X, Y) = \frac{Cov(X,Y)}{SD(X)\,SD(Y)}$$

This is estimated in practice using observed data:

$$\hat{\rho} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \times \sum_{i=1}^{n}(y_i-\bar{y})^2}} \equiv \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

- $|\hat{\rho}| \le 1$
- Values of $|\hat{\rho}|$ close to 1 indicate a strong linear relationship
- Values of $|\hat{\rho}|$ close to 0 indicate an absence of a linear relationship. } Strength
- Positive/Negative values of $\hat{\rho}$ signify positive/negative linear relationships } Direction

PGA Example: Data on $n = 196$ professional golfers

① y = driving accuracy, x = driving distance $\rightarrow \hat{\rho} = -0.61$

→ There is a reasonably strong negative linear relationship between x and y

→ As driving distance increases, driving accuracy decreases in a meaningful way.

② y = driving accuracy, x = putting accuracy $\rightarrow \hat{\rho} = 0.07$

→ There is a very weak positive linear relationship between x and y

→ Just because you're a good driver doesn't mean you're a good putter.

While the correlation coefficient describes the direction and strength of a linear relationship it does not tell us how much we can expect y to change for a given change in x. We also can't predict a value of y for a given change in x. For these things we need to fit a linear regression model.

Fitting a Simple Linear Regression Model (i.e., line of best fit)

Formally, for n pairs of observed values (x_i, y_i), $i = 1, 2, \ldots, n$ our model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Assumption: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ — iid: independent and identically distributed; σ^2: fixed constant

Consequence: $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Hence $E[y_i] = \beta_0 + \beta_1 x_i = \mu_i$ should on average be reasonably close to the true relationship between y and x. But we don't know μ_i, so need to first estimate it.

We call $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ the fitted value which is the estimated regression equation, and in the case of SLR, it's the equation of the line of best fit.

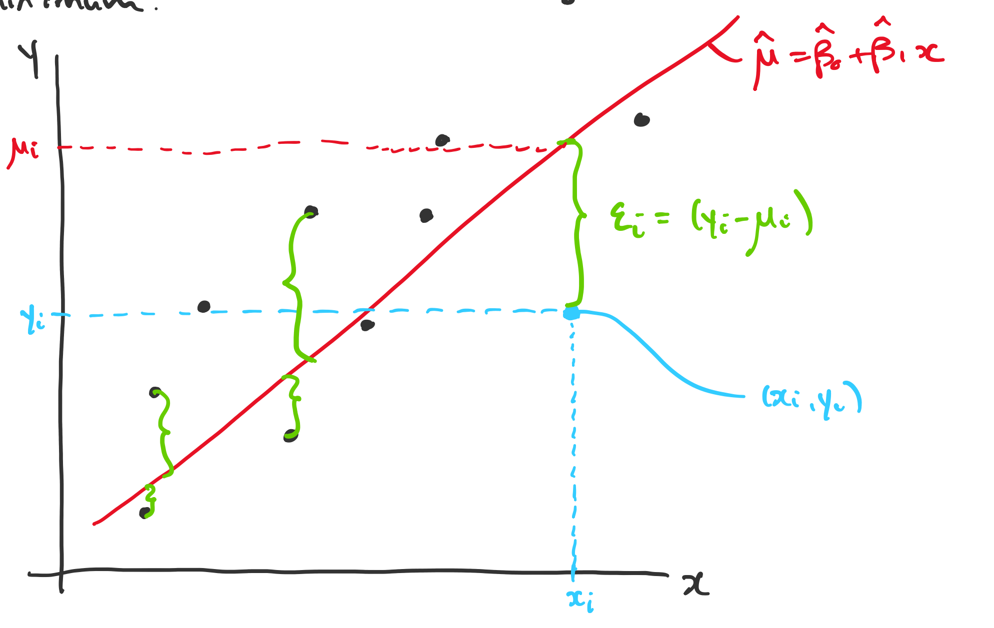
In order to obtain the estimates $\hat{\beta}_0, \hat{\beta}_1$ and hence $\hat{\mu}_i$, we can either use Maximum Likelihood Estimation (MLE) or Least Squares Estimation (LSE). Here, we consider LSE.

The idea behind LSE is that we want to obtain $\mu_i = \beta_0 + \beta_1 x_i$ that is "closest" to the points (x_i, y_i). As such the errors $\varepsilon_i = y_i - \mu_i = y_i - (\beta_0 + \beta_1 x_i)$ should be as small as possible.

In LSE we choose (β_0, β_1) such that

$$\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \mu_i)^2 = \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2 = S(\beta_0, \beta_1)$$

is at a minimum.



$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} 2[y_i - (\beta_0 + \beta_1 x_i)](-1) = -2\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]$$

$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} 2[y_i - (\beta_0 + \beta_1 x_i)](-x_i) = -2\sum_{i=1}^{n}x_i[y_i - (\beta_0 + \beta_1 x_i)]$$

Letting $\frac{\partial S}{\partial \beta_0} = 0$ and $\frac{\partial S}{\partial \beta_1} = 0$ and solving simultaneously yields:

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)] = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)$$

$$= n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\Rightarrow n\bar{y} = n\beta_0 + \beta_1 n\bar{x}$$

$$\Rightarrow \boxed{\beta_0 = \bar{y} - \beta_1 \bar{x}} \qquad (1)$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^{n} x_i[y_i - (\beta_0 + \beta_1 x_i)] = 0$$

Substituting (1):

$$\sum_{i=1}^{n} x_i[y_i - (\bar{y} - \beta_1\bar{x} + \beta_1 x_i)] = 0$$

$$\sum_{i=1}^{n} x_i[(y_i - \bar{y}) - \beta_1(x_i - \bar{x})] = 0$$

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) - \beta_1\sum_{i=1}^{n} x_i(x_i - \bar{x}) = 0$$

$$\boxed{\beta_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})}} \qquad (2)$$