

Part 2: A RM4Es summary of simple regression

SUMMARY: This Part covers technical details of simple regression model and discusses a research flow of building a simple regression model under the RM4Es framework.

In Part 1, we discussed simple regression from a structural point of view. In this Part, we will discuss simple regression from a process perspective and with more technical details.

As we remember, simple regression takes a representation where a dependent variable is expressed as a linear combination of a constant and an independent variable.

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- *Constant (intercept) β_0*
- *Slope β_1*
- *Residuals ε_i*

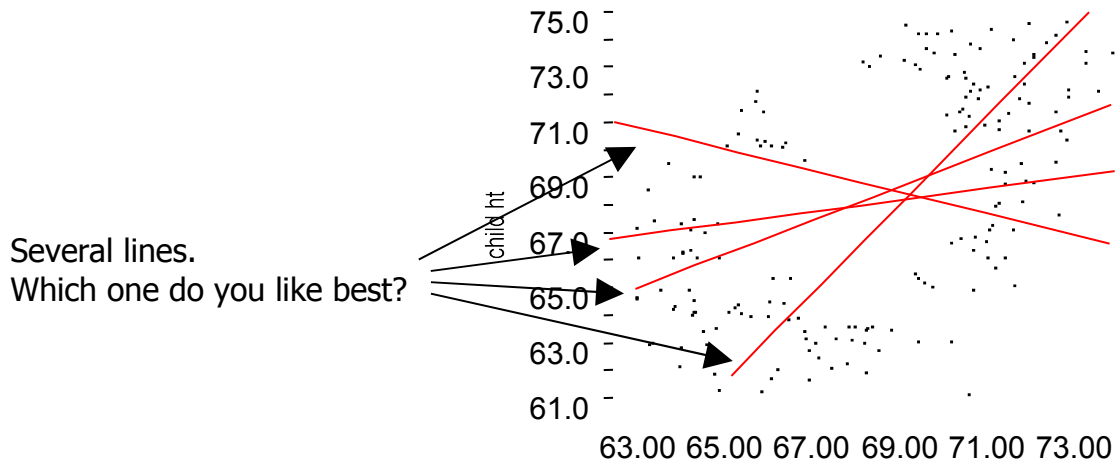
The goal of building a simple regression model is to find a constant and a slope that make our model a best representation of the reality we are studying.

In the first step of building a simple regression model, what we need to do is to consider what to be included in this model, which is called as model specification and equation specification in our language. In other words, we need to decide on what Y is and what X is. In practice, researchers often use subject knowledge, scatter plots of data or even guests to make decisions on this step.

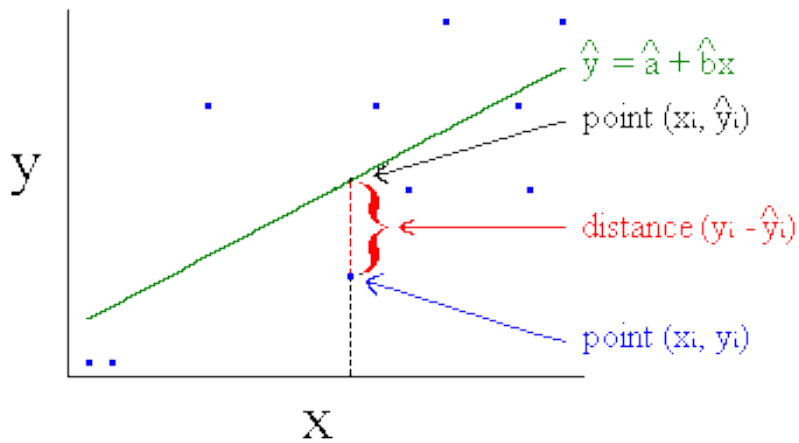
After Y and X are selected, we need to select β_0 and β_1 to complete this simple regression model. Simple regression assumes linear relationship between X and Y. Graphically, that means the relationship between X and Y can be represented by a straight line in a two dimension space. The values of β_0 and β_1 determine where the line is located, in this two dimension space. Therefore, the next step for our model building is to choose a line. Choosing a line is the same as choosing a slope (β_1) and an intercept (β_0).

As we know from our equation specification, each line we choose will determine the predictions of y values as based on x values. Those predictions come with errors, which are the differences between our predicted values and the related actual values. And quite often, we want the errors to be “small.” Therefore, finding a line making these errors small is quite a reasonable goal to achieve. That is called as a process to find a “best fit”.

- $\hat{Y}_i = \beta_0 + \beta_1 X_i$
- $r_i = Y_i - \hat{Y}_i$



The above figure illustrates what we discussed. There are several lines to choose. But which is the best? As illustrated below, for each line selected, there are associated errors $r_i = Y_i - \hat{Y}_i$ where Y_i is the actual value \hat{Y}_i is the predicted value or the value on the selected line. One way of finding a best fit is to minimize $\sum r_i^2$ that is called as ordinary least square method.



Specifically, for each observation in the data set, our selected line predicts where Y should be: $\hat{y}_j = b_0 + b_1 x_j$. The *residual* from j 'th data point is how far the true Y value is from where the line predicts $e_j = y_j - \hat{y}_j$

The sum of squared residuals (or sum of squared errors) gives an overall measure of how well the line fits: $SSE = e_1^2 + e_2^2 + \dots + e_n^2$. Then, we need to choose b_0 and b_1 to make SSE as small as possible.

To obtain calculation of β_0 and β_1 , we can follow the below steps.

- $RSS(\beta_0, \beta_1) = \sum (y_i - \beta_0 - \beta_1 x_i)^2$

Then, we differentiate RSS with respect to β_0 and β_1 , and set the derivatives equal to zero. We will obtain:

- $\beta_0 n + \beta_1 \sum x_i = \sum y_i$
- $\beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i$

Solve the equation will lead us to the following results:

- $b_1 = S_{XY} / S_{XX} = \sum (x_i - \text{mean}(x)) * (y_i - \text{mean}(y)) / \sum (x_i - \text{mean}(x))^2$

- $b_0 = \text{mean}(y) - b_1 \text{mean}(x)$

- $C_i = (x_i - \text{mean}(x)) / S_{XX}$
- $b_1 = \sum c_i y_i$

After we get our equation estimated, we would like to know how good our calculated b s are. In other words, we need to evaluate them.

- If $E(r_i) = 0$, $E(y_i) = b_0 + b_1 x_i$

Then,

- $E(b_0) = \beta_0$
- $E(b_1) = \beta_1$

So, we know that both are:

- Unbiased Estimator

- If $\text{cov}(e_i, e_j) = 0$, then $\text{var}(e_i) = \sigma^2$

- $\text{Var}(b_1) = \sigma^2 / \text{SXX}$
- $\text{Var}(b_0) = \sigma^2 [1/n + \text{Mean}(x)^2 / \text{SXX}]$

- Gauss-Markov Theorem: Smallest Possible of any linear unbiased estimation.

To sum, for the above discussion, we have used the following two assumptions:

- $E(r_i) = 0$

- $\text{cov}(e_i, e_j) = 0$

As we know, the regression line from the sample is not the regression line from the population. So we need to make inferences, that is, to explain our results.

What we want to do are to:

- Guess the slope of the population line.
- Guess what value Y would take for a given X value.

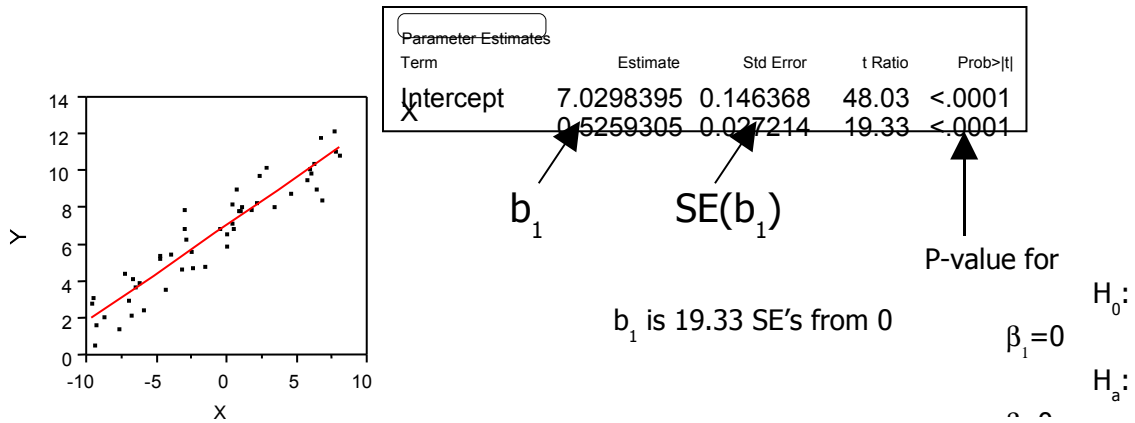
If r_i s are $\text{NID}(0, \sigma^2)$, b_0, b_1 are all normally distributed so we can construct confidence intervals.

- $[b_0 - t(0.05, n-2)\text{se}(b_0), b_0 + t(0.05, n-2)\text{se}(b_0)]$
- 95% Confidence Interval for the constant

- Construct 95% Confidence Interval for Slope

Here, r_i s are $\text{NID}(0, \sigma^2)$ is our assumption.

In words of hypothesis testing, small p-value for b_1 means there is a statistically detectable relationship between X and Y. Regression does a useful amount of explaining.



■ How to conduct hypothesis testing for B?

-
- B has a normal distribution

■ $B/se(b)$ has a T distribution

The assumption is still that r_i s are $NID(0, \sigma^2)$

When we are satisfied with our fitted model, we can explain our results to the users.

As we stated before, β_0 tells the value of Y when X is zero. And β_1 tells how much Y will change, when X changes by one unit.