LECTURE 3 Linear Modelling Multiple Regression

Objectives:

- Multiple Linear Regression
- Properties of Multiple Linear Regression
- R squared, Adjusted R squared,redicted R squared,
- Hypothesis Testing , Confidence interval for Multiple Linear Regression

Motivation for using Multiple Linear Regression

- The simple linear regression used only one regressor/explanatory variable. We had computed the R^2 which informs us regarding the variability in the response variable due to the explanatory variable. If suppose we are evaluating the level of Cholesterol with Age and suppose R^2 =50% then it implies that 50% variability in cholesterol is attributed to Age. How about rest of 50% of the variability? Part of it could be attributed to weight and diet. Therefore if we wish to explore linear relationship with multiple explanatory/ regressor variables then we use Multiple Linear Regression.

Multiple Linear Regression Model

Two Regressors/explanatory variables:

We need to estimate the population multiple linear regression relationship depicted by the general equation

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Specifically for the ith individual of the population

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

This informs us that there are 2 explanatory variables and this is the y value of the ith observation.

This can be estimated by our sample using the generic equation:

$$Y = A + B_1 X_1 + B_2 X_2 + E \quad \text{where}$$

$$\hat{Y} = A + B_1 X_1 + B_2 X_2$$

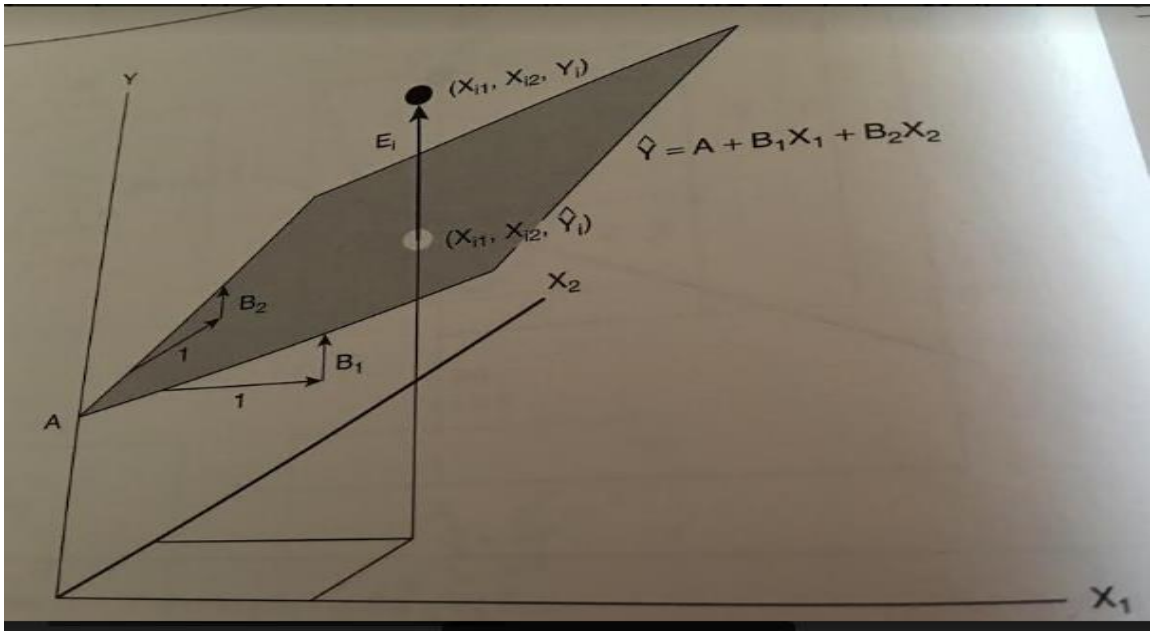
$$Y = \hat{Y} + E$$

Specifically for the ith individual of the sample

$$Y_i = A + B_1 X_{i1} + B_2 X_{i2} + E_i$$

$$E_i = Y_i - \hat{Y}_i = Y_i - (A + B_1 X_{i1} + B_2 X_{i2})$$

X_1, X_2 are two explanatory variables. The variables {X_1, X_2, Y} form a three dimensional space with the X_1 and X_2 forming a plane. This plane is the fitted plane.



The white dot is the Fitted point on the plane. Since the model does not pass through every real data point (observation) therefore there will be a residual.

$$E_i = Y_i - \hat{Y}_i = Y_i - (A + B_1 X_{i1} + B_2 X_{i2})$$

For the fitted plane to be as close as possible to the data values the residual sum of squared has to be minimized. Using this minimizing technique we calculate A and B (just like the Linear Regression).

$$S(A, B_1, B_2) = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y - A - B_i X_{i1} - B_2 X_{i2})^2$$

We will now differentiate this equation with respect to coefficient A, B_1 and B_2.

$$\frac{\partial S(A, B_1, B_2)}{\partial A} = \sum_{i=1}^{n} (-1)(2)(Y - A - B_i X_{i1} - B_2 X_{i2})$$

$$\frac{\partial S(A, B_1, B_2)}{\partial B_1} = \sum_{i=1}^{n} (-X_{i1})(2)(Y - A - B_i X_{i1} - B_2 X_{i2})$$

$$\frac{\partial S(A, B_1, B_2)}{\partial B_2} = \sum_{i=1}^{n} (-X_{i2})(2)(Y - A - B_i X_{i1} - B_2 X_{i2})$$

The normal equations obtained

$$An + B_1 \sum X_{i1} + B_2 \sum X_{i2} = \sum Y_i$$

$$A\sum X_{i1} + B_1\sum X_{i1}^2 + B_2\sum X_{i1}X_{i2} = \sum X_{i1}Y_i$$

$$A\sum X_{i2} + B_1\sum X_{i2}X_{i1} + B_2\sum X_{i2}^2 = \sum X_{i2}Y_i$$

After solving the three equations with three unknowns are:

Here $Y^* = Y_i - \overline{Y}$ $X^* = X_i - \overline{X}$ check

$$A = \overline{Y} - B_1\overline{X_1} - B_2\overline{X_2}$$

$$B_1 = \frac{\sum X_1^*Y^*\sum X_2^{*2} - \sum X_2^*Y^*\sum X_1^*X_2^*}{\sum X_1^{*2}\sum X_2^{*2} - (\sum X_1^*X_2^*)^2}$$

$$B_1 = \frac{\sum X_2^*Y^*\sum X_1^{*2} - \sum X_1^*Y^*\sum X_1^*X_2^*}{\sum X_1^{*2}\sum X_2^{*2} - (\sum X_1^*X_2^*)^2}$$

B_1 and B_2 are unique as long as the denominator is not zero.

$$\sum X_1^{*2}\sum X_2^{*2} \neq (\sum X_1^*X_2^*)^2$$

This condition is not true except when the variables perfectly correlated (collinear) or at least one of them is invariant ie at least one of them are not effected by any transformation applied to them.

Multiple Regression Interpretation:

The slope coefficient of the Multiple regression are called partial coefficients.

$$\hat{Y} = A + B_1X_1 + B_2X_2$$

Here B_1 informs us of the change of response variable $\hat{Y}$ with every unit of increase of X_1 keeping X_2 constant.

Here B_2 informs us of the change of response variable $\hat{Y}$ with every unit of increase of X_2 keeping X_1 constant.

Unlike the Multiple Linear Regression, Simple Linear Regression ignores the second explanatory variable. Multiple Linear regression keeps the second explanatory variable constant.

Multiple Regressor Variables:

$$\hat{Y} = A + B_1X_1 + B_2X_2 + B_3X_3 \ldots\ldots\ldots B_kX_k$$

For specific multi coordinate for one entity

$$\hat{Y}_i = A + B_1 X_{i1} + B_2 X_{i2} ..... B_k X_{ik}$$

$$Y = \hat{Y}_i + E_i$$

The normal equations can be formed but here these have a unique solution except the case when one of the explanatory variable is a linear combinations of the others or if one or more explanatory variables are invariant.

Residual Standard Error

Remember for the simple linear regression the residual standard error:

$$s_E^2 = (\frac{\sum_{i=1}^{n} E_i^2}{n-2})$$

This in case of multiple regression becomes:

$$s_E^2 = (\frac{\sum_{i=1}^{n} E_i^2}{n-k-1})$$

There are n-(k+1) degree of freedom for the residuals. All the Eis are not independent. The first n-(k+1) Eis can be chosen independently but the final k+1 have to chosen such that the k+1 constraints (k+1 normal equations) are satisfied.

$$1) E_1 + E_2 ..... + E_n = 0$$
$$2) E_1 X_1 + E_2 X_2 + ... E_n X_n = 0$$

.....

K+1) $E_2 X_2 + E_3 X_3 + ... E_k X_k = 0$ check

ANOVA TECHNIQUE

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} [(\hat{Y}_i - \bar{Y})]^2 + \sum_{i=1}^{n} [(Y_i - \hat{Y}_i)]^2$$

TSS	=	RegSS	+	RSS
Sum of squared total variation		Sum of Squared regression Explained By model		Sum of squared residual Not explained By the model

$R^2 = \frac{\text{Re}\, gSS}{TSS}$ Here the coefficient of determination informs us the variation of Y due to the multiple explanatory variable Xs.

The square root of R^2 gives the correlation coefficient which informs us regarding the correlation between the actual values of the response variable and the fitted values of response variables ie correlation between Y and $\hat{Y}$

Issues with R squared

a) The increase is number of explanatory variables might increase R squared though it might be just by chance due to chance correlations between explanatory variables. To adjust this artificial inflation Adjusted R squared is calculated by penalizing R squared.

As the number explanatory variables increase the squared multiple regression or coefficient of determination might increase maybe just by chance therefore to adjust for more predictors variables ie eliminate the increase in R squared which is due to the increase of predictors Adjusted R squared is used.

Mathematically

$$\tilde{R}^2 = 1 - \frac{S_E^2}{S_Y^2} = 1 - \frac{\frac{RSS}{n-k-1}}{\frac{TSS}{n-1}}$$

S_E is the Residual Standard Error for the Multiple Linear Regression.

S_Y is the sample standard deviation of Y

Adjusted R squared can be used to compare different models when there are different number of predictors.

Adjusted R squared is a unbiased predictor of the population R squared.

Situation: Suppose you add one explanatory variable at a time into the model. As you add these variables one at a time R squared and Adjusted R squared might increase. If you add another new explanatory variable and check the R squared and Adjusted R squared. If R squared increases but the Adjusted R squared does not increase then the new explanatory variable is extraneous and does not actually contribute to the model.

R squared or Adjusted R squared might not be a good predictor for new observations therefore predicted R squared is used. Predicted R squared checks how well the model fits the data. It checks if the model overfits the data. Predicted R squared removes each observation and estimating regression equation and then ascertaining how well this equation predicts the removed observation.

If R squared is higher(say 80%) than predicted R squared(50%) then it means that the model

Is overfitted.

Possibly the model has too many predictors and some should be dropped. The random noise/overfit is caused due to too many explanatory/regressor variables

Standardized Regression Coefficients

Social scientists sometimes wish to compare the effect of different explanatory variables. Suppose

We are measuring the an individuals Satisfaction level in life.

Satisfaction =dependent on income + Family_Relationships+ Work_Realtionships+ Education

These are measured on different scales. For instance income is measured in dollars wheras education is measured in years and the others are measured on a scale say 0-10. To standardize the scale we use the technique of Standard Normal. This technique is only effect if the distribution of the explanatory variables satisfies the assumption of normality. Rescaling does not change the relative slope of the coefficients.

Generally:

$$Y_i = A + B_1 X_{i1} + \ldots\ldots\ldots B_k X_{ik} + E_i \text{ -------1}$$

Also

$$\overline{Y} = A + B_1 \overline{X}_{i1} + \ldots\ldots\ldots B_k \overline{X}_{ik} + E_i \text{ -----2}$$

$$1-2 \Rightarrow Y_i - \overline{Y} = B_1 (X_{i1} - \overline{X}_{i1}) + \ldots\ldots\ldots B_k (X_{ik} - \overline{X}_{ik}) \text{ -----3}$$

Dividing by Standard deviation of Y as well as Multiplying and dividing the RHS equations by the SD of individual explanatory variables

$$\frac{Y_i - \overline{Y}}{S_Y} = (B_1 \frac{S_1}{S_Y}) \frac{(X_{i1} - \overline{X}_{i1})}{S_1} + \ldots\ldots\ldots (B_k \frac{S_k}{S_Y})(\frac{X_{ik} - \overline{X}_{ik}}{S_K}) + \frac{E_i}{S_Y} \text{ -------4}$$

$$Z_{iY} = B_i^* Z_{i1} + \ldots\ldots\ldots B_k^* Z_{ik} + E_i^*$$

The rescaling of the coefficients with respect to the standard deviations results in standardized regression coefficients enables the comparison of incommensurable effects of the explanatory variables.

Multiple Regression Results and:

The Multiple Regression model for the population is depicted as follows:

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} \ldots\ldots\ldots \ldots \beta_k x_{ik} + \varepsilon_i$$

The assumptions for this model(Multiple Linear Regression) are similar to Simple linear regression:

1) Linearity $E(\varepsilon_i) = 0$

2) Constant variance $V(\varepsilon_i) = \sigma_\varepsilon^2$
3) Normality $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$
4) Independence ε_i and ε_j are independent for i≠j
5) Fixed Xs

Under these assumptions the least square estimators A ,B_1,B_2……B_k of $\alpha, \beta_1, \ldots\ldots \beta_k$

are: unbiased, linear functions of the data, maximally efficient amongst unbiased estimators, maximum likelihood estimators, normally distributed.

Multiple Regression Sampling Variances

We are usually interested in the slope or the sampling variance of the slope.

The sampling variance of slope coefficient for multiple regression:

$$V(B_j) = \frac{1}{1-R_j^2} * \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x})^2} = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})^2}$$

R_j^2 Is the square multiple correlation, ie correlation from X_j on all others Xs. This will be high if there are is dependence of explanatory variable of x_j on other xs.

σ_ε^2 is smaller as compared to the case of simple linear regression because for the simple linear regression some of the errors are due to non-inclusive explanatory variables have been included in the model.

$\frac{1}{1-R_j^2}$ Is the variance inflation factor. Mathematically if R_j^2 is large then the Variance Inflation Factor is very high. Sometimes pairwise correlation might be small but multicolinearity might exist (linear dependence of three or variables on each other).

If the Variance Inflation Factor is high then multicollinearity exist.

VIF inflates the sampling variance of the slope coefficient.

If VIF=1 then there is no collinearity.

VIF=∞ If all the explanatory variables are correlated.

Confidence Interval and Hypothesis Test:

These are similar for Simple Linear Regression.

Standard Error of the slope coefficient B_j is as follows:

$$SE(B_j) = \frac{1}{\sqrt{1-R_j^2}} * \frac{s_E}{\sqrt{(x_{ij} - \bar{x})^2}}$$

Confidence interval is based on t distribution with n-k-1 degrees of freedom.

Hypothesis Testing for all Slopes

$$H_0 : \beta_1 = \beta_2 = ...\beta_k = 0$$

$$H_1 : At\ \text{least one of the}\ \beta_j \neq 0$$

If the explanatory variables are highly correlated then we will be able to reject H_o but we will not be able to reject individual hypothesis

$$H_0^{(1)} : \beta_1 = 0;\ H^{(2)}{}_0 : \beta_2 = 0 \ldots\ldots\ldots\ldots H^{(k)}{}_0 : \beta_2 = 0$$

F test for the main omnibus hypothesis is conducted by the F statistics

$$F_0 = \frac{\text{Re}\, gSS/k}{RSS/n-k-1} = \frac{n-k-1}{k} * \frac{R^2}{1-R^2}$$

The Total variation of Y can be partitioned by the ANOVA table:

Source	Sum of Squares	df	Mean Square	F
Regression	RegSS	k	RegSS/k	RegMS/RMS
Residuals	RSS	n-k-1	RSS/n-k-1	
Total	TSS	n-1		

For the Null Hypothesis to be true the ratio F=RegMS/RMS should be as close to 1 as possible.

$$R^2 = 1 - \frac{RSS}{TSS} \qquad 0 \leq R^2 \leq 1$$

R^2=1 if RSS=0 This will be possible if the fitted model explains 100% of the variability in Y.

If the RegMS is much larger than RMS then we Reject H_0.

Sometimes instead of taking all the slopes into consideration some slopes are taken. This is usually when quantitative variables are taken into consideration.