# 1 Multiple Linear Regression

- Previously we have looked at Simple Linear Regression - the case of one dependent variable Y explained by **one** independent variable X.
- Multiple regression analysis is an extension of simple regression analysis, as described previously, to applications involving the use of two or more independent variables (predictors) to estimate the value of the dependent variable (response variable).
- In the case of two independent variables, denoted by X1 and X2, the linear algebraic model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- The definitions of the above terms are equivalent to the definitions in previous classes for simple regression analysis, except that more than one independent variable is involved in the present case.
- Based on sample data, the linear regression equation for the case of two independent variables is

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

- We refer to $b_0, b_1, b_2 \ldots$ as ***regression coefficients***. These coefficients are estimate for parameter values $\beta_0, \beta_1, \beta_2 \ldots$.
  *No estimation carried out for the random sampling error term $\epsilon$.*
- The multiple regression equation identifies the best-fitting line based on the method of ***Ordinary Least Squares***. In the case of multiple regression analysis, the best-fitting line is a line through n-dimensional space (3-dimensional in the case of two independent variables).

- **Important:** we will denote the number of predictor variables (a.k.a independent variables) as $p$. Some resources uses $k$. (Be Familiar with Both).
- The calculations required for determining the values of the parameter estimates in a multiple regression equation and the associated standard error values are quite complex and generally involve matrix algebra. However, computer software, such as `R`, is widely available for carrying out such calculations.

## 1.1 Statistical Assumptions

The assumptions of multiple linear regression analysis are similar to those of the simple case involving only one independent variable. For point estimation, the principal assumptions are that

(1) the dependent variable is a continuous random variable ,

(2) the relationship between the several independent variables and the one dependent variable is *linear* (as opposed to quadratic or cubic - this is something we will explore more later).

Additional assumptions for statistical inference (estimation or hypothesis testing) are that

(3) the variances of the conditional distributions of the dependent variable, given various combinations of values of the independent variables, are all equal (*homoscedascity - something we will look at in a forthcoming lecture*),

(4) the conditional distributions of the dependent variable are normally distributed (*i.e. Residuals are nomally distributed*),

(5) the observed values of the dependent variable are independent of each other. (*Violation of this assumption is called autocorrelation. - Again, we will return to this later.*)

## 1.2 Implementing a MLR model using R

Implementing a MLR model in `R` is quite simple and very similar to fitting an SLR model. All one has to do is to specify the additional predictor variables, using the following structure:

```
myModel = lm(Y ~ X1 + X2 + .....)
```

## 1.3 Example: Cheese Tasting

- As an example, we shall use data on the taste of cheese, suggested in *Introduction to the Practice of Statistics by D.S. Moore and G.P. McCabe, (Freeman, 1998).*
- The data give scores for the taste of a cheese (`Taste`) from 30 different formulations which caused variation in the concentration in the cheese of *acetic acid* (`Acetic`), *hydrogen sulphide* (`H2S`) and *lactic acid* (`Lactic`).
- One would wish to model the dependence of the taste score on the concentrations of those three constituents, using the thirty observations.

```
> FitAll

Call:
lm(formula = Taste ~ Acetic + H2S + Lactic, data = Cheese)

Coefficients:
(Intercept)       Acetic          H2S       Lactic
   -28.8768       0.3277       3.9118      19.6705
```

The fitted model is therefore (using 2 decimal places)

$$TasteEstimate = -28.87 + 0.33Acetic + 3.91H2S + 19.67Lactic$$

Remark: It is acceptable ( in fact preferred ) to write as follows:

$$\hat{Y} = -28.87 + 0.33\ X1 + 3.91\ X2 + 19.67\ X3$$

while stating that Y refers the dependent variable taste and $X_1$,$X_2$ and $X_3$ refer to the three independent variables. (Remember to state which is which).

```
> FitAll = lm(Taste ~ Acetic + H2S + Lactic, data = Cheese)
> summary(FitAll)

Call:
lm(formula = Taste ~ Acetic + H2S + Lactic, data = Cheese)

Residuals:
    Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
Acetic        0.3277     4.4598   0.073  0.94198
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared: 0.6518,     Adjusted R-squared: 0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

## 2 The Coefficient of Determination

The coefficient of determination, denoted $R^2$ and and pronounced ***R squared***, is a number that indicates how well data fit a statistical model, sometimes simply a line or a curve. It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, as the proportion of total variation of outcomes explained by the model.

### Formula

$R^2$ is the proportion of variance in Y explained by a linear function of X.

$$R^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}$$

- In the case of Simple Linear Regression only, the Coefficient of Determination has the same value as the square of the Pearson Correlation Coefficient.
- If the Pearson Correlation Coefficent is 0.8, then the Coefficient of Determination is $0.8^2 = 0.64$
- However this is not the case in Multiple Linear Regression. Hence we are not putting too much emphasis on the relationship between the two measures.

**Important:** The Coefficient of Determination equation can be expressed in term of ***Sums of Squares Identities*** that also appear in the regression ANOVA table.

$$R^2 = \frac{SS_{\text{reg}}}{SS_{\text{tot}}} = \left(1 - \frac{SS_{\text{err}}}{SS_{\text{tot}}}\right).$$

- We will use the names "The coefficient of determination", "$R^2$" and "R squared" interchangably.
- **Important** - The coefficient of determination, $R^2$, is a measure of the proportion of variability explained by, or due to the **linear relationship** in a sample of data.
- **Important** - $R^2$ is a number between zero and one .

  $$0 \leq R^2 \leq 1$$

  A value close to zero suggests a poor model. A value close to 1 indicates an excellent model
- **Important** - A very high value of $R^2$ can arise even though the relationship between the dependent and independent variables is **non-linear**. The fit of a model should never simply be judged from the $R^2$ value. It is advisable to construct a scatterplot to visually assess the relationship.
- In the case of simple linear regression only (i.e. bivariate data) the coefficient of determination is equivalent to the square of the correlation coefficient of X and Y.
- The $R^2$ value is presented as part of the output of the `summary()` command for a fitted model. In the R code output, it is referred to as "`multiple R square`".
- *There is also adjusted R square, which is going to be a significant part of the MA4505 syllabus.(Model Selection and Overfitting etc)*
- **Important** : If given the variance of the sample the dependent variable, while knowing the sample size $n$ - you can quickly compute $SS_{tot}$. Furthermore if you are given a value for $R^2$, you can compute $SS_{reg}$. This will enable you to construct the Regression ANOVA table.

### Code Output

- The coefficient of determination is listed as "`Multiple R-squared`" in a `summary` output.
- Also given on this output is the F Test statistic for the ANOVA table and the corresponding p-value.

```
> summary(lm(Abs2 ~Conc))

...........

Residual standard error: 0.007026 on 5 degrees of freedom
Multiple R-squared:  0.9994,    Adjusted R-squared:  0.9993
F-statistic:  8980 on 1 and 5 DF,  p-value: 2.481e-09
............
```