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 ϵ.
- 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 not going to be a major part of the MA4605 syllabus.*
- **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
............
```

3 ANOVA for Regression

An ANOVA-F test can be constructed to test overall (global) fit of the linear regression model.

- In the ANOVA procedure, a hypothesis test (known as an F test) is used to test for the significance of the **overall model**.
- That is, it is used to test the null hypothesis that there is no relationship in the population between the (several) independent variables taken as a group and the one dependent variable.
- Specifically, the null hypothesis states that all of the coefficients in the regression equation for the population are equal to zero.
- Therefore, for the case of several independent variables, or predictors, the null hypothesis with respect to the F test is

H0: $E[Y] = \beta_0$

H1: $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$

- Under the Null Hypothesis $\beta_1 = \beta_2 = \ldots = 0$.
- The import is that the predictor variables **collectively** are not useful in predicting values for Y (i.e. computing an expected value for Y- denoted $E[Y]$).
- The Alternative Hypothesis expresses the case the using the predictor variables help the predictive capability of the model.

The ANOVA Table for Linear Regression

- This table is for both the simple and multiple linear regression cases. For simple linear regression, the number of predictor variables is 1.
- SS_{reg} is the regression sum of squares, also called the explained sum of squares
- SS_{tot} is the total sum of squares (proportional to the variance of the data)
- $SS_{\text{reg}} = \sum_i (\hat{Y}_i - \bar{Y})^2$
- $SS_{\text{tot}} = \sum_i (Y_i - \bar{Y})^2$
- n is the number of obversations.
- p (or sometimes k) is the number of predictor variables.

Source	DF	Sum of Squares	Mean Square	F
Regression	p	SS_{reg}	$MS_{\text{reg}} = SS_{\text{reg}}/\text{p}$	$\text{F} = MS_{\text{reg}}/MS_{\text{err}}$
Error	n-p-1	SS_{err}	$MS_{\text{err}} = SS_{\text{err}}/(\text{n-p-1})$	
Total	n-1	SS_{tot}		

Remark

- SS_{tot} is related to the sample variance of the response variable as follows:

$$\text{var}(Y) = \frac{SS_{\text{tot}}}{n-1}$$

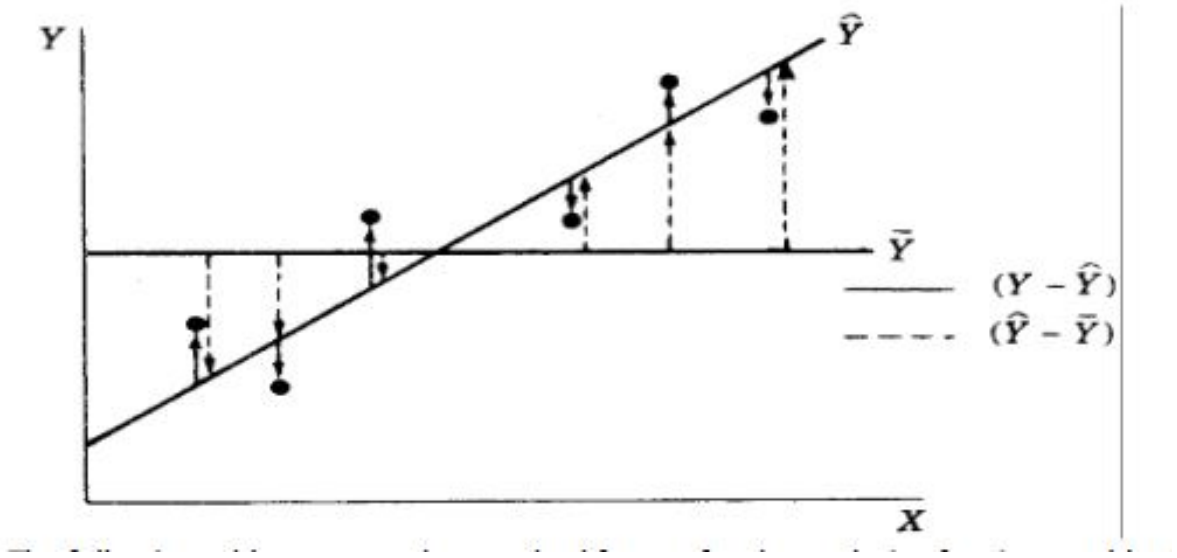
- In an exam situation, you can expect to be given the sample standard deviation of the response variable
- For example, in the Cheese Tast example, the standard deviation of the response variable **Taste** is given below. You can determine the sample variance from it.

```
> sd(Taste)
[1] 16.25538
> var(Taste)
[1] 264.2375
```

- Remark: When there is only one independent variable in the regression model, then the ANOVA procedure is equivalent to a two-tail t-test directed at the slope ($H0 : \beta_1 = 0$). Therefore, use of the ANOVA procedure is often not required in simple regression analysis in practice, as it doesnt not provide additional information.

3.1 Going into More Detail: Separating Variances

- The ANOVA method helps us to choose the best way of plotting a curve from amongst the many that are available. Analysis of variance (ANOVA) provides such a method in all cases where we maintain the assumption that the errors occur only in the y-direction.
- In such situations there are two sources of y-direction variation in a regression model fit.
    1. The first is the **variation due to regression**, i.e. due to the relationship between the variables.
    2. The second is the random experimental error in the y-values, which is called the **variation about regression**.



- ANOVA is a powerful method for separating two sources of variation in such situations

- **Important:** If there is no regression effect in the population, then the fitted value (sloped) line differs from the mean (horizontal) line purely by chance. It follows that the variance estimate based on the differences - called mean square regression (MSR), would be different only by chance from the variance estimate based on the residuals called mean square error (MSE).
- On the other hand, if there is a regression effect, then the mean square regression is inflated in value as compared with the mean square error. the regression line significantly differs from the mean (horizontal) line.
- The following table presents the standard format for the analysis of variance table that is used to test for the significance of an overall regression effect.
- The degrees of freedom k (also denoted p) associated with MSR in the table is the number of independent variables in the multiple regression equation.
- As indicated in the table, the test statistic is the ratio of the two values. The p-value for the test statistic is provided in code output.

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F ratio
Regression (R)	k	SSR	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Sampling error (E)	$n-k-1$	SSE	$MSE = \frac{SSE}{n-k-1}$	
Total (T)	$n-1$	SST		

The ANOVA table can be obtained for the regression model with the `anova()` command, when the model is specified. Using a data set from last weeks lab:

```
> anova(FitA)
Analysis of Variance Table

Response: Abso
          Df  Sum Sq Mean Sq F value    Pr(>F)
Conc       1 0.44327 0.44327  8979.5 2.481e-09 ***
Residuals  5 0.00025 0.00005
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4 Sample Question (Q6d Autumn 2008/09 Dr. N Coffey)

Q6. A hospital administrator wished to study the relation between patient satisfaction (Y) and the patient's age (X_1, in years). She randomly selected 22 patients and collected the data some of which is presented below, where larger values of Y indicated more satisfaction.

i	1	2	3	4	5	6	7	8	⋯	22
y_i	48	57	66	70	89	36	46	54	⋯	52
x_{i1}	50	36	40	41	28	49	42	45	⋯	44

The MINITAB printout for fitting the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

is as follows:

```
Regression Analysis: Satisfaction versus Age, Severity of , Anxiety Leve

The regression equation is
Satisfaction = 163 - 1.21 Age - 0.666 Severity of Illness - 8.6 Anxiety Level

Predictor              Coef  SE Coef
Constant             162.47    26.51
Age                 -1.2179   0.3112
Severity of Illness -0.6505   0.8452
Anxiety Level         -8.69    12.56

S = 10.2895

Analysis of Variance

Source          DF       SS      MS     F
Regression      (i)  4137.2  1379.1  (vi)
Residual Error (ii)     (iv)     (v)
Total         (iii)  6143.3
```

(i) Complete the ANOVA table by filling in the values for (i)-(vi). Conduct a hypothesis test to determine the significance of the linear regression model.

- Sample size n = 22
- For MA4605 2015, you would be told that the sample standard deviation for the response variable Y. For this question : $S_Y = 17.1033$
- Number of predictor variables $p = 3$
- $df_1 = p = 3$
- $df_2 = n - p - 1 = 18$
- In general, you will be give the p-values for various possible Test Statistics. All you have to do is select the correct one, and interpret it as not-significant/significant etc.
- Compute the correct Test Statistic, and determine the corresponding $p-$value from a table provided. For this exercise the p-value is 0.000124.
- In some cases, such as the examples to follow, The $p-$value will be given in the code. Youl will just be asked to verify the corresponding Test Statistic.
- Express your conclusions on whether or not model is useful from a predictive point of view (i.e regression coefficients are jointly significant).
- From the previous example - can you compute the value for "Multiple R Squared"? (0.67 approx)
- From the previous example - is it possible to compute Pearson Correlation coefficients **directly**.

5 The Cheese Taste Data

Summary Outputs from this Week's lab (Lab D) exercises. (Note: 30 observations)

```
> summary(Fit1)
....
Call:
lm(formula = Taste ~ Acetic + H2S
....
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -26.940     21.194  -1.271 0.214536
Acetic         3.801      4.505   0.844 0.406245
H2S            5.146      1.209   4.255 0.000225 ***
...
Residual standard error: 10.89 on 27 degrees of freedom
Multiple R-squared:  0.5822,    Adjusted R-squared:  0.5512
F-statistic: 18.81 on 2 and 27 DF,  p-value: 7.645e-06
```

```
> summary(Fit2)
....
Call:
lm(formula = Taste ~ Acetic + Lactic)
....
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -51.366     21.174  -2.426  0.02223 *
Acetic         5.571      4.761   1.170  0.25217
Lactic        31.392      8.956   3.505  0.00161 **
....
Residual standard error: 11.67 on 27 degrees of freedom
Multiple R-squared:  0.5203,    Adjusted R-squared:  0.4847
F-statistic: 14.64 on 2 and 27 DF,  p-value: 4.936e-05
```

```
> summary(Fit3)
Call:
lm(formula = Taste ~ H2S + Lactic)
....
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.592      8.982  -3.072  0.00481 **
H2S            3.946      1.136   3.475  0.00174 **
Lactic        19.887      7.959   2.499  0.01885 *
---
....
Residual standard error: 9.942 on 27 degrees of freedom
Multiple R-squared:  0.6517,    Adjusted R-squared:  0.6259
F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07
```

```
> summary(FitAll)

Call:
lm(formula = Taste ~ Acetic + H2S + Lactic)
....
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 *
....
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
```