3. Multiple Linear Regression

The multiple linear regression refers to a model where you have *k* independent variables (predictors). One of the objectives of the analysis might be to evaluate the significance of each of the *k* independent variables and how they relate to the dependent variable ***Y***.

The multiple linear regression model is not restricted to modelling only planes. By using higher-ordered terms, such as quadratic or cubic powers of the ***X'***s or cross products of one ***X*** with another, more complex surfaces can be modelled.

Statistical Model

In the multiple regression setting, because of the potentially large number of predictors, it is more efficient to use matrices to define the regression model and the subsequent analyses.

We start with the simple case first. Consider the following simple linear regression function:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad i = 1, \ldots, n$$

Letting $i = 1, \ldots, n$ in fact means the following *n* equations:

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 x_1 + \varepsilon_1 \\ y_2 &= \beta_1 + \beta_2 x_2 + \varepsilon_2 \\ &\cdots \\ y_n &= \beta_1 + \beta_2 x_n + \varepsilon_n \end{aligned}$$

These equations can be formulated in matrix notation as:

$$\underset{n\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} = \underset{(n\times 2)}{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}} \underset{(2\times 1)}{\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}} + \underset{n\times 1}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

This formulation can be generalised for the multiple regression model. So the population regression model is:

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \qquad i = 1, \ldots, n$$

which can be expressed in terms of four vectors/matrices:

$$\underset{(n \times 1)}{\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}, \qquad \underset{(n \times k)}{\boldsymbol{X} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix}}, \qquad \underset{(k \times 1)}{\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}}, \qquad \underset{(n \times 1)}{\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}$$

$\boldsymbol{y}$: the $\boldsymbol{n} \times 1$ column vector of values of the dependent variable

$\boldsymbol{X}$: the $\boldsymbol{n} \times \boldsymbol{k}$ matrix consisting of a column of ones, followed by the $\boldsymbol{k} - 1$ column vectors of the values of the independent variables.

$\boldsymbol{\beta}$: the $\boldsymbol{k} \times 1$ vector of parameters to be estimated

$\boldsymbol{\varepsilon}$: the $\boldsymbol{n} \times 1$ vector of random errors

Hence, the linear model may now be written in matrix notation as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Classical Assumptions

1. The independent variables are fixed.
2. The rank of $\boldsymbol{X}$ is equal to k.
3. $E(\boldsymbol{\varepsilon}) = 0, \qquad Var(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon\varepsilon'}) = \sigma^2 \boldsymbol{I}_n = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \vdots & 0 & \ldots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}.$

The least squares (LS) estimator of $\boldsymbol{\beta}$,**b,** is obtained by minimising the sum of square function:

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{y'y} - 2\boldsymbol{y'X\beta} + \boldsymbol{\beta'X'X\beta}$$

$$\frac{\partial S}{\partial \beta} = -2\boldsymbol{X'y} + 2\boldsymbol{X'X\beta} = 0$$

$$\frac{\partial^2 S}{\partial \beta \partial \beta'} = 2\boldsymbol{X'X}, \text{ p.d.}$$

In matrix notation, the normal equations are written as

$$\boldsymbol{X'Xb = X'y}$$

If $\boldsymbol{X'X}$ has an inverse, then the normal equations will have the unique solution

$$\boldsymbol{b = (X'X)^{-1} X'y}$$

Properties of the LS estimator

The estimator $\boldsymbol{b}$ can be written as $\boldsymbol{b = \beta + (X'X)^{-1} X'\varepsilon}$

1. $E(\boldsymbol{b}) = \boldsymbol{\beta}$
2. $Var(\boldsymbol{b}) = \sigma^2 \boldsymbol{(X'X)^{-1}}$
3. If the error is multi-normal, so is the LS estimator: $\boldsymbol{b} \sim \boldsymbol{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{(X'X)^{-1}})$

Gauss-Markov Theorem

In the class of all *linear*, *unbiased* estimators, LS estimator has the minimum variance, ie, it is **BLUE**. That is to say any other linear and unbiased estimator has a variance-covariance matrix which exceeds that of $\boldsymbol{b}$ by a p.s.d matrix

Residuals

As in the case of the simple regression, the residuals are defined by the difference between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$:

$$\boldsymbol{e = y - \hat{y} = y - Xb = y - X(X'X)^{-1} X'y = My}$$

where, $\boldsymbol{M = I - X(X'X)^{-1} X'}$, is a symmetric, idempotent matrix, ie. $\boldsymbol{M = M' = MM = M'M}$.

Assuming that $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I})$ it follows that $\boldsymbol{b} \sim \boldsymbol{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{(X'X)^{-1}})$.

The variance-covariance matrix of $\boldsymbol{b}$ is in fact

$$\begin{bmatrix} \mathrm{var}(b_1) & Cov(b_1,b_2) & \dots & Cov(b_1,b_k) \\ Cov(b_2,b_1) & \mathrm{var}(b_2) & \dots & Cov(b_2,b_k) \\ \dots & \dots & \dots & \dots \\ Cov(b_k,b_1) & Cov(b_k,b_2) & \dots & \mathrm{var}(b_k) \end{bmatrix}$$

The constant variance of this matrix, σ^2, is estimated by $s^2 = \frac{\boldsymbol{e'e}}{n-k}$.

It can also be shown that $\frac{\boldsymbol{\varepsilon'M\varepsilon}}{\sigma^2} \sim \chi^2_{n-k}$, where ($n$-$k$) is the trace of $\boldsymbol{M}$. Since $\boldsymbol{e} = \boldsymbol{M\varepsilon}$ and $\boldsymbol{e'e} = \boldsymbol{\varepsilon'M\varepsilon}$, then it follows that $\frac{\boldsymbol{e'e}}{\sigma^2} \sim \chi^2_{n-k}$ and consequently $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$.

These results are the basis of testing hypotheses of regression and coefficients.

Tests on Individual Coefficient

Consider $\boldsymbol{b} \sim \boldsymbol{N}(\boldsymbol{\beta}, \sigma^2 \boldsymbol{(X'X)^{-1}})$, then the individual coefficient is distributed as $b_j \sim N(\beta_j, \sigma^2 c_{jj}), (j = 1,...,k)$, where c_{jj} is the j^{th} diagonal element of $\boldsymbol{(X'X)^{-1}}$. Recalling the definition of the *t*-distribution, we have:

$$\frac{b_j - \beta_j}{\sigma\sqrt{c_{jj}}} \sim N(0,1)$$

$$\frac{\frac{b_j - \beta_j}{\sigma\sqrt{c_{jj}}}}{\sqrt{\frac{(n-k)s^2}{\sigma^2}/(n-k)}} = \frac{b_j - \beta_j}{s\sqrt{c_{jj}}} \sim t_{n-k} \qquad \left(\frac{(b_j - \beta_j)^2}{s^2 c_{jj}} \sim F(1, n-k)\right)$$

Note that the null hypothesis of significance for an individual coefficient $H_0 : \beta_j = 0$ will reduce the statistic to

$$\frac{b_j}{s\sqrt{c_{jj}}} = \frac{b_j}{SE(b_j)} \sim t_{n-k}.$$

Analysis of Variance

The table of ANOVA for the multiple linear model is defined as:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares
	SS	df	MS
Regression	$\boldsymbol{b'X'y} - \frac{1}{n}\boldsymbol{y'i\,i'y}$	k-1	RSS/k-1
Error	$\boldsymbol{y'y}$ - $\boldsymbol{b'X'y}$	n-k	ESS/(n-k)
Total	$\boldsymbol{y'y} - \frac{1}{n}\boldsymbol{y'i\,i'y}$	n-1	

This can be used to calculate the *coefficient of determination* R^2.

$$R^2 = \frac{\boldsymbol{b'X'y} - \frac{1}{n}\boldsymbol{y'i\,i'y}}{\boldsymbol{y'y} - \frac{1}{n}\boldsymbol{y'i\,i'y}}$$

R^2 will always be increased by adding more independent variables, therefore, in multiple regression an *adjusted* version of it is used:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

Testing Overall Significance of Regression

The overall (joint) significance of slope coefficients $H_0 : \beta_2 = ... = \beta_k = 0$ is tested by the results provided in ANOVA table. Under the null hypothesis:

$$F = \frac{MSR}{MSE} = \frac{(\boldsymbol{b'X'y} - \frac{1}{n}\boldsymbol{y'i\,i'y})/(n-1)}{(\boldsymbol{y'y}\text{ - }\boldsymbol{b'X'y})/(n-k)} = \frac{R^2/(n-1)}{(1-R^2)/(n-k)}$$

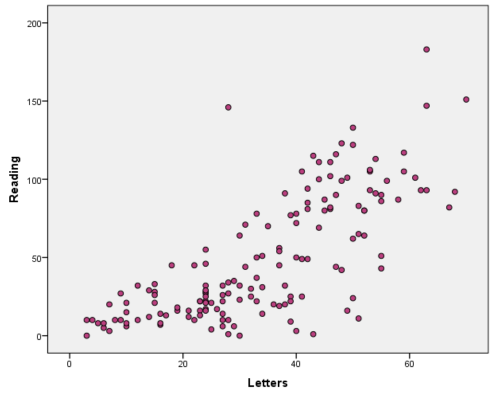
Example 3.1

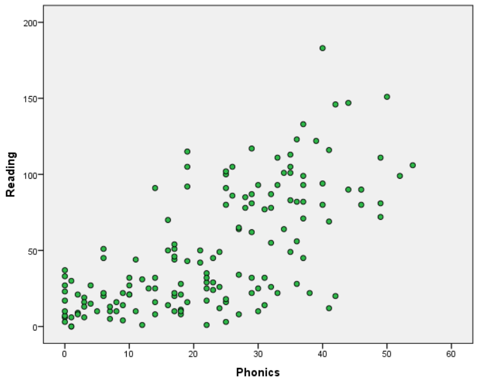
Data were collected by a school district, in the US, to assess the reading skill progress of students in their first year of formal schooling. A simple random sample of students was selected from all the students in the district. The variables of interest in the data set are:

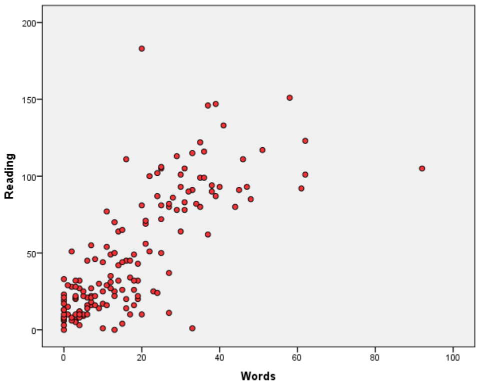
Letters	score on letter identification test;
Phonics	score on letter sound test;
Words	score on word identification test;
Reading	score on reading test after a year.

The school district is interested to see if the reading skill progress of children is determined by their test results.

First, we examine the relationship between **Reading** and **Words**, **Letters** and **Phonics**.







All three predictors appear to have a positive relationship with response. It appears that **Words** has the least variability in its distribution and **Phonics** has the greatest variability.

The population regression model is:

$$\textbf{Reading} = \beta_1 + \beta_2\,\textbf{Letters} + \beta_3\ \textbf{Phonics} + \ \beta_4\,\textbf{Words} + \varepsilon$$

We also inspect the descriptives.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Reading	179	0	183	47.76	39.812
Letters	163	3	72	34.10	15.980
Phonics	163	0	54	22.21	13.743
Words	163	0	92	18.03	16.009
Valid N (listwise)	154				

Now, consider the results of Model Summary and ANOVA and comment.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.835[a]	.696	.690	22.129

a. Predictors: (Constant), Words, Phonics, Letters

ANOVA [a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	168542.977	3	56180.992	114.728	.000[b]
	Residual	73453.133	150	489.688		
	Total	241996.110	153			

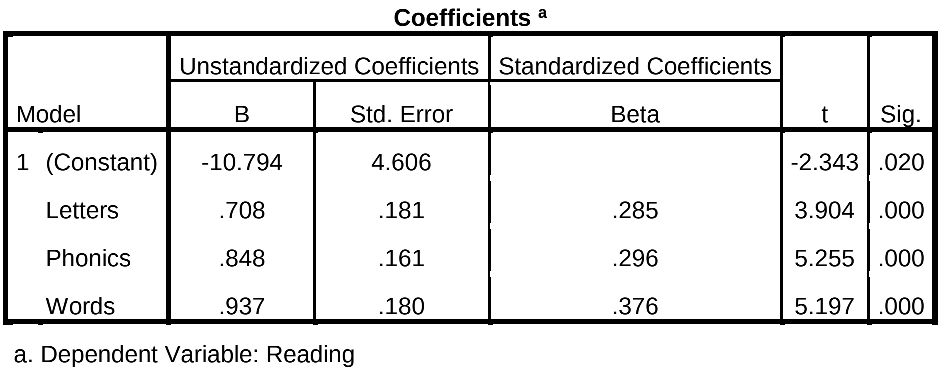
a. Dependent Variable: Reading

b. Predictors: (Constant), Words, Phonics, Letters

The Regression Summary table indicates that the R-Square is 0.696. However, the R-Square always increases as you include more terms in the model and, as a result, can be misleading.

The Adjusted R-Square is a measure similar to R-Square but takes into account the number of terms in the model. Therefore, when you compare models, it is more appropriate to compare the Adjusted R-Square. The Adjusted R-Square for this model is 0.690. The percentage of the variability in the response variable that is explained by the model is approximately 69%.

The overall significance of coefficients $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ is tested by (F = 114.728) the p-value for the model which is less than 0.001. Because this is smaller than any reasonable alpha level, you reject the null hypothesis and conclude that at least one slope is not equal to zero.



Coefficients [a]

Model	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
1 (Constant)	-10.794	4.606		-2.343	.020
Letters	.708	.181	.285	3.904	.000
Phonics	.848	.161	.296	5.255	.000
Words	.937	.180	.376	5.197	.000

a. Dependent Variable: Reading

The estimated (fitted) regression equation is:

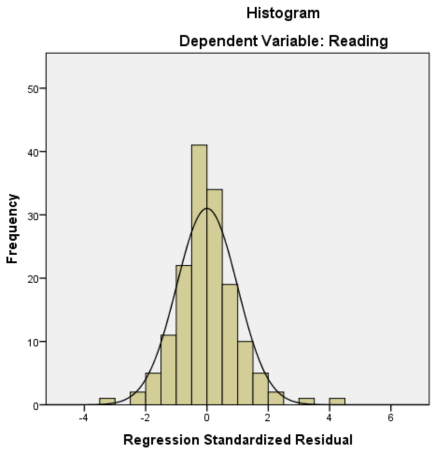
Reading = –10.794 + 0.708***Letters** + 0.848***Phonics**+ 0.937***Words**

The *t*-values and *p*-values in the Coefficients table test the null hypothesis that the slope for each of the independent variables is equal to zero, $H_0 : \beta_j = 0, (j = 1, ..., k)$. Looking at these values, and presuming an alpha equal to 0.05, you reject the null hypothesis in each case. All of these variables are significant in predicting/determining **Reading**.

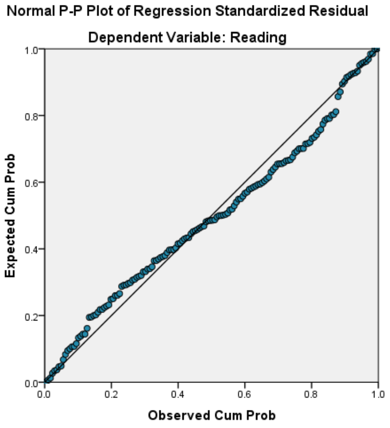
Note: You should be careful when interpreting the tests of hypothesis for the parameter estimates. They test the significance of each variable when it is added to a model that contains all of the other independent variables. As a result, if the independent variables in the model are correlated with one another, the significance of both variables can be hidden in these tests. Therefore, you should not remove more than one variable at a time from the model, based on these tests.

Brief Diagnostics

As with any statistical analysis, the assumptions of the analysis should be examined to ensure they were met.


Histogram

Dependent Variable: Reading


Normal P-P Plot of Regression Standardized Residual

Dependent Variable: Reading

Descriptives

			Statistic	Std. Error
Standardized Residual	Mean		.0000000	.07978836
	95% Confidence Interval for Mean	Lower Bound	-.1576291	
		Upper Bound	.1576291	
	5% Trimmed Mean		-.0153725	
	Median		-.0361754	
	Variance		.980	
	Std. Deviation		.99014754	
	Minimum		-3.08328	
	Maximum		4.36254	
	Range		7.44582	
	Interquartile Range		1.03231	
	Skewness		.513	.195
	Kurtosis		2.684	.389

The skewness statistic is 0.513. The mean is bigger than the median, which indicates a slight skewed-to-the-right distribution of the residuals.

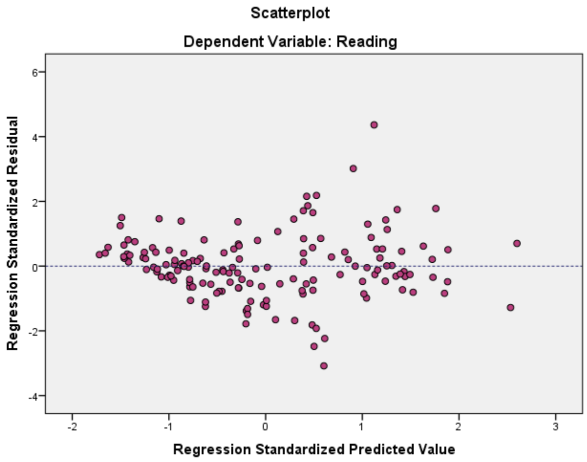
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.077	154	.025	.965	154	.001

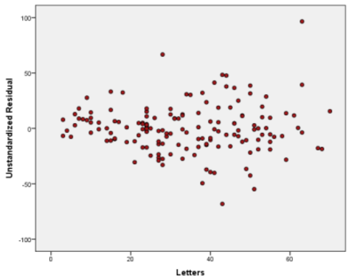
a. Lilliefors Significance Correction

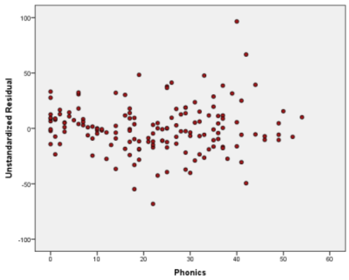
The normality tests reject the null hypothesis of normally distributed residuals. These results should be evaluated in conjunction with the histogram and normal probability plots of the residuals. All in all, there does not seem to be serious problems about Normality assumption.

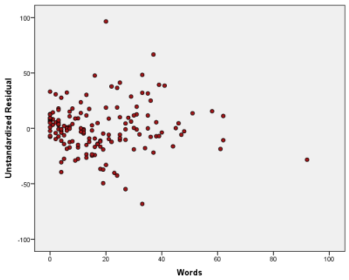
Note: All the normality tests depend on the sample size. As the sample size becomes larger, increasingly smaller departures from normality can be detected. Small sample size will likely yield a less powerful test and you might want to use a higher alpha value.



Although there is a slight pattern of expanding variation, but it is not severe to indicate any problem with homoscedasticity assumption. In fact the following plots show that this pattern is mostly caused by the variable **Words**.







Practical Week 3

The variables the **Bodyfat.sav** data set are measurements of percentage of body fat, age, weight, height and other physical measurements recorded for 252 men by Roger W Johnson of Calvin College in Minnesota. The following variables are in the data set:

Case	Case Number
PctBodyFat	Percent body fat
Density	Density (gm/cm^3)
Age	Age (yrs)
Weight	Weight (lbs)
Height	Height (inches)
Adioposity	Adiposity index
FatFreeWt	Fat Free Weight
Neck	Neck circumference (cm)
Chest	Chest circumference (cm)
Abdomen	Abdomen circumference (cm)
Hip	Hip circumference (cm)
Thigh	Thigh circumference (cm)
Knee	Knee circumference (cm)
Ankle	Ankle circumference (cm)
Biceps	Extended biceps circumference (cm)
Forearm	Forearm circumference (cm)
Wrist	Wrist circumference (cm)

- Run a regression of **PctBodyFat** on the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm** and **Wrist**.
- Explore the existence of linear relations between the dependent variable and independent variables by appropriate graphs.
- Comment on relevant results of Model Summary, ANOVA and Coefficients tables with clear interpretations of tests and their conclusions.
- Reduce the model by eliminating insignificant variables one by one. Note the changes in the *p*-value for the model, the R^2 and adjusted R^2 and the parameter estimates and their *p*-values.
- Write down the final estimated equation and comment.