MULTIPLE LINEAR REGRESSION ANALYSIS USING MICROSOFT EXCEL

by Michael L. Orlov
Chemistry Department, Oregon State University (1996)

INTRODUCTION

In modern science, regression analysis is a necessary part of virtually almost any data reduction process. Popular spreadsheet programs, such as Quattro Pro, Microsoft Excel, and Lotus 1-2-3 provide comprehensive statistical program packages, which include a regression tool among many others.

Usually, the regression module is explained clearly enough in on-line help and spreadsheet documentation (i.e. items in the regression *input* dialog box). However, the description of the *output* is minimal and is often a mystery for the user who is unfamiliar with certain statistical concepts.

The objective of this short handout is to give a more detailed description of the regression tool and to touch upon related statistical topics in a hopefully readable manner. It is designed for science undergraduate and graduate students inexperienced in statistical matters. The regression output in Microsoft Excel is pretty standard and is chosen as a basis for illustrations and examples (Quattro Pro and Lotus 1-2-3 use an almost identical format).

CLASSIFICATION OF REGRESSION MODELS

In a regression analysis we study the relationship, called **the regression function**, between one variable **y**, called the **dependent variable**, and several others $\mathbf{x_i}$, called the **independent variables**. Regression function also involves a set of unknown parameters $\mathbf{b_i}$. If a regression function is linear in the parameters (*but not necessarily in the independent variables !*) we term it a **linear regression model**. Otherwise, the model is called **non-linear**. Linear regression models with more than one independent variable are referred to as **multiple linear models,** as opposed to **simple linear models** with one independent variable.

The following notation is used in this work:

y	- dependent variable (*predicted by a regression model*)
y^*	- dependent variable (*experimental value*)
p	- number of independent variables (number of coefficients)
x_i $(i=1,2, \dots p)$	- ith independent variable from total set of p variables
b_i $(i=1,2, \dots p)$	- ith coefficient corresponding to x_i
b_0	- intercept (or constant)
$k=p+1$	- total number of parameters including intercept (constant)
n	- number of observations (experimental data points)
$i =1,2 \dots p$	- independent variables' index
$j=1,2, \dots n$	- data points' index

Now let us illustrate the classification of regression models with mathematical expressions:

Multiple linear model

General formula:

$$\mathbf{y = b_0 + b_1x_1 + b_2x_2 + \dots b_px_p} \tag{1}$$

or

$$\mathbf{y = b_0 + \Sigma_i\, b_ix_i \quad i=1,2,\dots p} \tag{1a}$$

Polynomial (model is linear in parameters , but not in independent variables):

$y = b_0 + b_1x + b_2x^2 + b_3x^3 \dots b_px^p$, which is just a specific case of (1)

with $x_1 = x, x_2 = x^2, x_3 = x^3 \dots\dots x_p = x^p$

Simple linear model

$$y = b_0 + b_1x_1$$

It is obvious that simple linear model is just specific case of multiple one with k=2 (p=1)

Non-linear model

$y = A(1-e^{-Bx})$,
where A, B are parameters

In further discussion we restrict ourselves to multiple linear regression analysis.

MAIN OBJECTIVES OF MULTIPLE LINEAR REGRESSION ANALYSIS

Our primary goal is to determine the best set of parameters $\mathbf{b_i}$, such that the model predicts experimental values of the dependent variable as accurately as possible (i.e. calculated values $\mathbf{y_j}$ should be close to experimental values $\mathbf{y_j*}$).

We also wish to judge whether our model itself is adequate to fit the observed experimental data (i.e. whether we chose the correct mathematical form of it).

We need to check whether all terms in our model are significant (i.e. is the improvement in "goodness" of fit due to the addition of a certain term to the model bigger than the noise in experimental data).

DESCRIPTION OF REGRESSION INPUT AND OUTPUT

The standard regression output of spreadsheet programs provides information to reach the objectives raised in the previous section. Now we explain how to do that and touch upon related statistical terms and definitions.

The following numerical *example* will be used throughout the handout to illustrate the discussion:

Table 1. Original experimental data

Data point # j	y*	z
1	20.6947	2.5
2	28.5623	3.1
3	157.0020	8.1
4	334.6340	12.2
5	406.5697	13.5
6	696.0331	17.9
7	945.1385	21.0

We choose **y*** to be the dependent experimental observable and **z** to be the independent one. Suppose we have, say, theoretical reasons to believe that relationship between two is:

$$\mathbf{y* = b_0 + b_1*z + b_2*z^2 + b_3*z^3}$$

We can rewrite this expression in form (1):

$$\mathbf{y = b_0 + b_1x_1 + b_2x_2 + b_3x_3,} \text{ where} \qquad \mathbf{(1b)}$$
$$\mathbf{x_1 = z, x_2 = z^2} \text{ and } \mathbf{x_2 = z^3}$$

In the next step we prepare the spreadsheet input table for regression analysis:

Table 2. Regression input

Data point #	**Dependent var.**	**Independent variables**		
j	y*	x_1(=z)	x_2(=z^2)	x_3(=z^3)
1	20.6947	2.5	6.25	15.63
2	28.5623	3.1	9.61	29.79
3	157.0020	8.1	65.61	531.44
4	334.6340	12.2	148.84	1815.85
5	406.5697	13.5	182.25	2460.38
6	696.0331	17.9	320.41	5735.34
7	945.1385	21.0	441.00	9261.00

In order to perform a regression analysis we choose from the Microsoft Excel menu*:

Tools ⟶ Data analysis ⟶ Regression

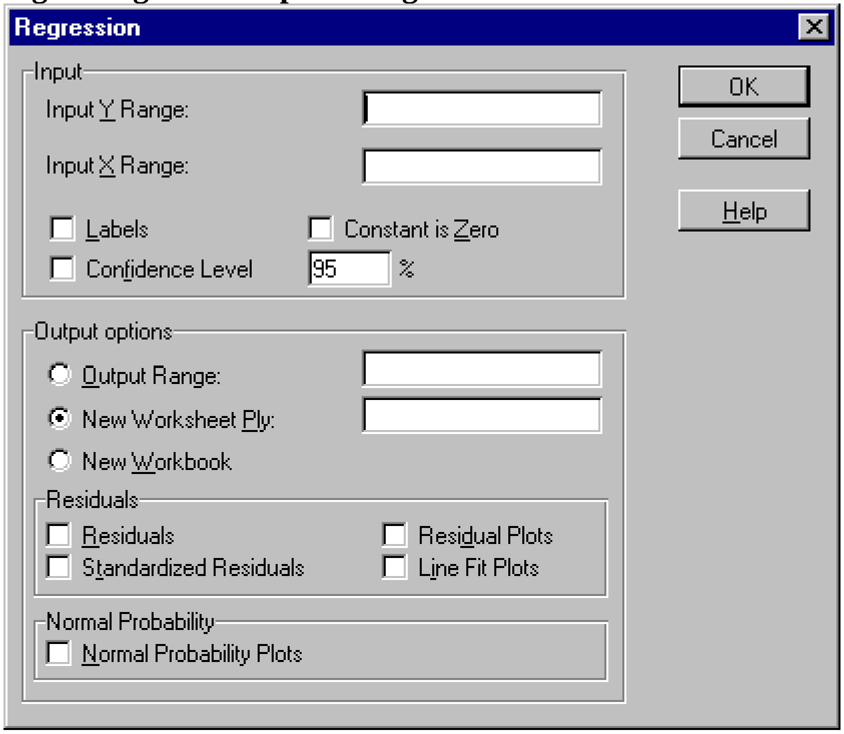
Note that data analysis tool should have been previously added to Microsoft Excel during the program setup (Tools – Add-Ins – Analysis ToolPak).

The pop-up input dialog box is shown on Fig.1. Elements of this box are described in on-line help. Most of them become clear in the course of our discussion as well.

The **"Input Y range"** refers to the spreadsheet cells containing the independent variable **y*** and the **"Input X range"** to those containing independent variables **x** (in our example $\mathbf{x = x_1, x_2, x_3}$**)** (see Table 2). If we do not want to force our model through the origin we leave the **"Constant is Zero"** box unchecked. The meaning of **"Confidence level"** entry will become clear later. The block **"Output options"** allows one to choose the content and locations of the regression output. The minimal output has two parts *"Regression Statistics"* and *"ANOVA*" (ANalysis Of VAriance). Checking the appropriate boxes in subblocks *"Residuals"* and *"Normal Probability"* will expand the default output information. We omit from our discussion description of *"Normal Probability" output.* Now we are ready to proceed with the discussion of the regression output.

* - *In **Quattro Pro** the sequence is **Tools - Numeric Tools - Analysis Tools - Advanced Regression**. In the last step instead of "Advanced Regression", one can choose "Regression" from the menu. In this case the simplified regression output will be obtained.*

Fig. 1. Regression input dialog box



Regression

Input

Input Y Range:

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

<u>Residual output</u>

<u>Example</u>

In Microsoft Excel the residual output has the following format:

Table3. Residual output*

Observation (j)	**Predicted Y (y_j)**	**Residuals (r)**	**Standard Residuals (r')**
1	20.4424	0.2523	0.3351
2	28.9772	-0.4149	-0.5511
3	156.3982	0.6038	0.8020
4	335.5517	-0.9178	-1.2189
5	406.3355	0.2342	0.3111
6	695.6173	0.4159	0.5524
7	945.3121	-0.1736	-0.2305

** - Corresponding notation used in this handout is given in parenthesis*

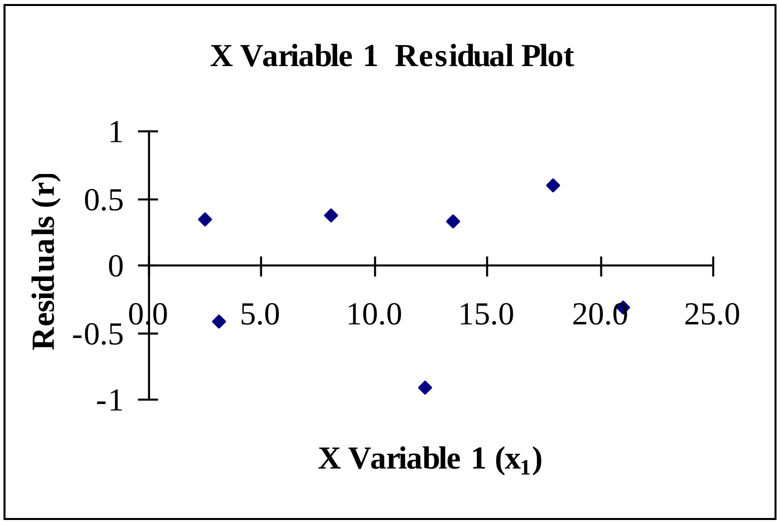
Residual (or **error,** or **deviation***)* is the difference between the observed value **y*** of the dependent variable for the jth experimental data point (x_{1j}, x_{2j}, …, x_{pj}, y_j*) and the corresponding value $\mathbf{y_j}$ given by the regression function $y_j = b_0 + b_1x_{1j} + b_2x_{2j} + \dots b_px_{pj}$ ($y_j = b_0 + b_1x_{1j} + b_2x_{2j} + b_3x_{3j}$ in our example):

$$\mathbf{r_j = y_j^* - y_j} \qquad \mathbf{(2)}$$

Parameters **b** (b_0, b_1, b_2, … b_p) are part of the *ANOVA* output (discussed later).

If there is an obvious correlation between the residuals and the independent variable **x** (say, residuals systematically increase with increasing **x**)**,** it means that the chosen model is not adequate to fit the experiment (e.g. we may need to add an extra term $\mathbf{x_4=z^4}$ to our model (1b)). A plot of residuals is very helpful in detecting such a correlation. This plot will be included in the regression output if the box *"Residual Plots"* was checked in the regression input dialog window (Fig. 1).

<u>*Example*</u>



However, the fact that the residuals look random and that there is no obvious correlation with the variable **x** does not necessarily mean by itself that the model is adequate. More tests are needed.

Standard (or **standardized**) **residual** is a residual scaled with respect to the *standard error (deviation*) $\mathbf{S_y}$ in a dependent variable:

$$\mathbf{r_j' = r_j / S_y} \qquad \mathbf{(2a)}$$

The quantity $\mathbf{S_y}$ is part of the *"Regression statistics"* output (discussed later). Standardized residuals are used for some statistical tests, which are not usually needed for models in physical sciences.

ANOVA output

There are two tables in ANOVA (Analysis of Variance).

Example

Table 4. ANOVA output (part I)*

	df	**SS**	**MS**	**F**	**Significance F**
Regression	3 (df_R)	723630.06 (SS_R)	241210.02 (MS_R)	425507.02 (F_R)	6.12E-09 (P_R)
Residual (error)	3 (df_E)	1.70 (SS_E)	0.57 (MS_E)		
Total	6 (df_T)	723631.76 (SS_T)	N/A (MS_T)		

Table 4a. ANOVA output (part II)*

	Coefficients (b_i)	**Standard Error** (se (b_i))	**t Stat** (t_i)	**P-value** (P_i)	**Lower 95%** ($b_{L,(1-Pi)}$)	**Upper 95%** ($b_{U,(1-Pi)}$)
Intercept (b_0)	0.52292226	1.77984111	0.293802778	0.7881	-5.1413318	6.1871763
X Variable 1 (x_1)	2.91437225	0.73039587	3.990126957	0.0282	0.58992443	5.2388201
X Variable 2 (x_2)	2.02376459	0.07318737	27.65182747	0.0001	1.79084949	2.2566797
X Variable 3 (x_3)	-0.0009602	0.00206174	-0.46574477	0.6731	-0.0075216	0.0056011

** - Corresponding notation used in this handout is given in parenthesis; N/A means "not available" in Microsoft Excel regression output.*

Coefficients.
The regression program determines the best set of parameters **b** (b_0, b_1, b_2, … b_p) in the model $y_j=b_0 +b_1x_{1j}+b_2x_{2j}+\dots b_px_{pj}$ by minimizing the *error sum of squares* SS_E (discussed later). Coefficients are listed in the second table of ANOVA (see Table 4a). These coefficients allow the program to calculate *predicted* values of the dependent variable **y** (y_1, y_2, … y_n), which were used above in formula (2) and are part of *Residual output* (Table 3).

Sum of squares.
In general, the sum of squares of some arbitrary variable **q** is determined as:

$$\mathbf{SS_q = \Sigma_j{}^n(q_j - q_{avg})^2}\text{, where} \qquad \mathbf{(3)}$$

$\mathbf{q_j}$ - jth observation out of **n** total observations of quantity **q**
$\mathbf{q_{avg}}$ - average value of **q** in **n** observations: $\mathbf{q_{avg}= (\Sigma_j{}^n\, q_j)/n}$

In the ANOVA regression output one will find three types of sum of squares (see Table 4):

***1). Total sum of squares* $\mathbf{SS_T}$:**

$$\mathbf{SS_T = \Sigma_j^n(y_j^* - y^*_{avg})^2}, \text{ where} \qquad \mathbf{(3a)}$$
$$\mathbf{y^*_{avg} = (\Sigma_j^n\ y_j^*)/n}$$

It is obvious that $\mathbf{SS_T}$ is the sum of squares of deviations of the experimental values of dependent variable **y*** from its average value. $\mathbf{SS_T}$ could be interpreted as the sum of deviations of **y*** from the simplest possible model (**y** is constant and does not depend on any variable **x**):

$$\mathbf{y = b_0}, \qquad \text{with } \mathbf{b_0} = \mathbf{y^*_{avg}} \qquad \mathbf{(4)}$$

$\mathbf{SS_T}$ has two contributors: **residual (error) sum of squares ($\mathbf{SS_E}$)** and **regression sum of squares($\mathbf{SS_R}$)**:

$$\mathbf{SS_T = SS_E + SS_R} \qquad \mathbf{(5)}$$

***2). Residual* (*or* *error) sum of squares* $\mathbf{SS_E}$:**

$$\mathbf{SS_E = \Sigma_j^n\ (r_j - r_{avg})^2} \qquad \mathbf{(6)}$$

Since in the underlying theory the *expected* value of residuals $\mathbf{r_{avg}}$ is assumed to be zero, expression (6) simplifies to:

$$\mathbf{SS_E = \Sigma_j^n\ (r_j)^2} \qquad \mathbf{(6a)}$$

The significance of this quantity is that by the minimization of $\mathbf{SS_E}$ the spreadsheet regression tool determines the best set of parameters **b=** b_0, b_1, b_2, …, b_p for a given regression model. $\mathbf{SS_E}$ could be also viewed as the due-to-random-scattering-of -**y***-about-predicted-line contributor to the total sum of squares $\mathbf{SS_T}$. This is the reason for calling the quantity “due to *error (residual)* sum of squares”.

***3). Regression sum of squares* $\mathbf{SS_R}$:**

$$\mathbf{SS_R = \Sigma_j^n\ (y_j - y^*_{avg})^2} \qquad \mathbf{(7)}$$

$\mathbf{SS_R}$ is the sum of squares of deviations of the *predicted-by-regression-model* values of dependent variable **y** from its average *experimental* value $\mathbf{y^*_{avg}}$. It accounts for addition of **p** variables (x_1, x_2, …, x_p) to the simplest possible model (4) (variable **y** is just a constant and does not depend on variables **x**), **i.e.** $\mathbf{y = b_0}$ **vs.** $\mathbf{y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p}$. Since this is a transformation from the “*non-regression* model” (4) to the true *regression* model (1), $\mathbf{SS_R}$ is called the “due to *regression* sum of squares”.

The definition of $\mathbf{SS_R}$ in the form (7) is not always given in the literature. One can find different expressions in books on statistics [1, 2] :

$$\mathbf{SS_R = \Sigma_i^p b_i \Sigma_j^n (x_{i\,j} - x_{avg}) y_j,} \text{ where} \qquad \mathbf{(7a)}$$

$$\mathbf{x_{avg} = (\Sigma_j^n x_j{*})/n}$$

or

$$\mathbf{SS_R = \Sigma_i^p b_i \Sigma_j^n x_{i\,j} y_j{*} - (\Sigma_j^n y_j{*})^2 /n} \qquad \mathbf{(7b)}$$

Relationships (7a-b) give the same numerical result, however, it is difficult to see the physical meaning of $\mathbf{SS_R}$ from them**.**

Mean square (variance) and degrees of freedom

The general expression for the mean square of an arbitrary quantity **q** is:

$$\mathbf{MS_q = SS_q / df} \qquad \mathbf{(8)}$$

$\mathbf{SS_q}$ is defined by (3) and **df** is the **number of degrees of freedom** associated with quantity $\mathbf{SS_q}$. **MS** is also often referred to as the **variance.** The number of degrees of freedom could be viewed as the difference between the number of observations **n** and the number of constraints (fixed parameters associated with the corresponding sum of squares $\mathbf{SS_q}$).

1). Total mean square $\mathrm{MS_T}$ ***(total variance)*****:**

$$\mathbf{MS_T = SS_T/(n - 1)} \qquad \mathbf{(9)}$$

$\mathbf{SS_T}$ is associated with the model (4), which has only one constraint (parameter $\mathbf{b_0}$), therefore the number of degrees of freedom in this case is:

$$\mathbf{df_T = n - 1} \qquad \mathbf{(10)}$$

2). Residual (error) mean square $\mathrm{MS_E}$ ***(error variance)*****:**

$$\mathbf{MS_E = SS_E / (n - k)} \qquad \mathbf{(11)}$$

$\mathbf{SS_E}$ is associated with the random error around the regression model (1), which has **k=p+1** parameters (one per each variable out of **p** variables total plus intercept). It means there are **k** constraints and the number of degrees of freedom is :

$$\mathbf{df_E = n - k} \qquad \mathbf{(12)}$$

***3). Regression mean square* $\mathbf{MS_R}$ (*regression variance*):**

$$\mathbf{MS_R = SS_R /(k - 1)} \tag{13}$$

The number of degrees of freedom in this case can be viewed as the difference between the total number of degrees of freedom $\mathbf{df_T}$ (10) and the number of degrees of freedom for residuals $\mathbf{df_E}$ (12) :

$$\mathbf{df_R = df_T - df_E = (n - 1) - (n - k)}$$

$$\mathbf{df_R = k - 1 = p} \tag{14}$$

Tests of significance and F-numbers

The F-number is the quantity which can be used to test for the statistical difference between two variances. For example, if we have two random variables **q** and **v**, the corresponding F- number is:

$$\mathbf{F_{qv} = MS_q / MS_v} \tag{15}$$

The variances $\mathbf{MS_q}$ and $\mathbf{MS_v}$ are defined by an expression of type (8). In order to tell whether two variances are statistically different, we determine the corresponding probability **P** from F-distribution function:

$$\mathbf{P{=}P(F_{qv}, df_q, df_v)} \tag{16}$$

The quantities $\mathbf{df_q, df_v}$ - degrees of freedom for numerator and denominator - are parameters of this function. Tabulated numerical values of **P** for the F-distribution can be found in various texts on statistics or simply determined in a spreadsheet directly by using the corresponding statistical function (e.g. in Microsoft Excel one would use $\mathbf{FDIST(F_{qv}, df_q, df_v)}$ to return the numerical value of P). An interested reader can find the analytical form of $\mathbf{P{=}P(F_{qv}, df_q, df_v)}$ in the literature (e.g. [1, p.383]).

The probability **P** given by (16) is a probability that the variances $\mathbf{MS_q}$ and $\mathbf{MS_v}$ are statistically *indistinguishable*. On the other hand, **1-P** is the probability that they are *different* and is often called **confidence level.** Conventionally, a reasonable confidence level is 0.95 or higher. If it turns out that $\mathbf{1\text{-}P < 0.95}$, we say that $\mathbf{MS_q}$ and $\mathbf{MS_v}$ are statistically the same. If $\mathbf{1\text{-}P > 0.95}$, we say that at least with the 0.95 (or 95%) confidence $\mathbf{MS_q}$ and $\mathbf{MS_v}$ are different. The higher the confidence level, the more reliable our conclusion. The procedure just described is called the **F-test**.

There are several F-tests related to regression analysis. We will discuss the three most common ones. They deal with significance of parameters in the regression model . The first and the last of them is performed by spreadsheet regression tool automatically, whereas the second one is not.

1). Significance test of all coefficients in the regression model

In this case we ask ourselves: ”With what level of confidence can we state that AT LEAST ONE of the *coefficients* **b** (b_1, b_2, … b_p) in the regression model is significantly different from zero?”. The first step is to calculate the F-number for the whole regression (part of the regression output (see Table 4)):

$$\mathbf{F_R = MS_R / MS_E} \qquad \textbf{(17)}$$

The second step is to determine the numerical value of the corresponding probability $\mathbf{P_R}$ (also part of the regression output (see Table 4)) :

$$\mathbf{P_R = FDIST(F_R, df_R, df_E)} \qquad \textbf{(18)}$$

Taking into account expressions (12) and (14) we obtain:

$$\mathbf{P_R = FDIST(F_R, k - 1, n - k)} \qquad \textbf{(18a)}$$

Finally we can determine the confidence level $\mathbf{1 - P_R}$. At this level of confidence, the variance “due to regression” $\mathbf{MS_R}$ is statistically different from the variance “due to error” $\mathbf{MS_E}$. In its turn it means that the addition of **p** variables (x_1, x_2, …, x_p) to the simplest model (4) (dependent variable **y** is just a constant) is a statistically significant improvement of the fit. Thus, at the confidence level not less than $\mathbf{1- P_R}$ we can say: “***At least ONE of coefficients in the model is significant***”. $\mathbf{F_R}$ could be also used to compare two models describing the same experimental data: the higher $\mathbf{F_R}$ the more adequate the corresponding model.

<u>*Example*</u>

In our illustrative exercise we have $\mathbf{P_R}$**= 6.12E-09** (Table 4), the corresponding level of confidence $\mathbf{1 - P_R}$ = **0.9999**. Therefore with the confidence close to 100% we can say that at *least one* of coefficients b_1, b_2 and b_3 is significant for the model $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$**,** where x_1=z, x_2= z^2 and x_3= z^3.

NOTE: From this test, however, we **<u>can not</u>** be sure that ***ALL coefficients b_1 ,b_2 and b_3 are non-zero .***

If $\mathbf{1- P_R}$ is not big enough (usually less than 0.95), we conclude that ***ALL the coefficients in the regression model are zero*** (in other words, the hypothesis that “the variable **y** is just a constant” is better than “it is function of variables **x** (x_1, x_2, …, x_p) **”**).

2). Significance test of subset of coefficients in the regression model

Now we want to decide "With what level of confidence can we be sure that at least ONE of the coefficients in a selected subset of all the coefficients is significant?". Let us test a subset of the last **m** coefficients in the model with a total of **p** coefficients (b_1, b_2, … b_p). Here we need to consider two models:

$$\mathbf{y = b_0 + b_1x_1 + b_2x_2 + \ldots\ b_px_p} \qquad \text{(unrestricted)} \qquad \mathbf{(19)}$$

and

$$\mathbf{y = b'_0 + b'_1x_1 + b'_2x + \ldots\ b'_{p-m}x_{p-m}} \qquad \text{(restricted)} \qquad \mathbf{(20)}$$

These models are called **unrestricted** (19) and **restricted** (20) respectively. We need to perform *two separate* least square regression analyses for each model.

From the regression output (see Table 4) for each model we obtain the corresponding error sum of squares $\mathbf{SS_E}$ and $\mathbf{SS'_E}$ as well as variance $\mathbf{MS_E}$ for the unrestricted model. The next step is to calculate the F-number for testing a subset of **m** variables "by hand" (it is not part of Microsoft Excel ANOVA for an obvious reason, i.e. *you* must decide how many variables to include in the subset):

$$\mathbf{F_m = \{(SS'_E - SS_E) / m\} / MS_E} \qquad \mathbf{(21)}$$

$\mathbf{F_m}$ could be viewed as an indicator of whether the reduction in the error variance due to the addition of the subset of **m** variables to the restricted model (20) (**($\mathbf{SS'_E - SS_E}$) / m**) is statistically significant with respect to the overall error variance $\mathbf{MS_E}$ for the unrestricted model (19). It is equivalent to testing the hypothesis that at least one of coefficients in the subset is not zero. In the final step, we determine probability $\mathbf{P_m}$ (also "by hand")**:**

$$\mathbf{P_m = FDIST(F_m, m, n - k)} \qquad \mathbf{(22)}$$

At the confidence level **1- $\mathbf{P_m}$** ***at least ONE of the coefficients in the subset of m is significant***. If **1- $\mathbf{P_m}$** is not big enough (less than 0.95) we state that ***ALL m coefficients in the subset are insignificant***.

Example

The regression output for the unrestricted model ($y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$**,** where x_1=z, x_2= z^2 and x_3= z^3) is presented in Table 4. Say, we want to test whether the quadratic and the cubic terms are significant. In this case the restricted model is:

$$\mathbf{y = b_0 + b_1x_1,} \quad \text{(restricted model)} \quad \mathbf{(23)}$$
$$\mathbf{where\ x_1 = z}$$

The subset of parameters consists of two parameter and **m=2**. By analogy with the input table for the unrestricted model (Table 2) we prepare one for the restricted model:

Table 5. Regression input for restricted model

Data point # j	**Dependent var.** y*	**Independent var.** x_1(=z)
1	20.6947	2.5
2	28.5623	3.1
3	157.0020	8.1
4	334.6340	12.2
5	406.5697	13.5
6	696.0331	17.9
7	945.1385	21.0

We perform an additional regression using this input table and as part of ANOVA obtain:

Table 6. Regression ANOVA output for the restricted model

	df'	**SS'**	**MS'**	**F'**	**Significance F'**
Regression	1	689216	689216	100	1.70-4
Residual (error)	5	***34415.70***	6883		
Total	6	723632			

From Table 4 and Table 6 we have:

$\mathbf{SS_E = 1.70}$ (error sum of squares; unrestricted model)
$\mathbf{MS_E = 0.57}$ (error mean square; unrestricted model)
$\mathbf{df_E = (n - k) = 3}$ (degrees of freedom; unrestricted model)
$\mathbf{SS'_E = 34415.70}$ (error sum of squares; restricted model)

Now we are able to calculate $\mathbf{F_{m=2}}$:

$$\mathbf{F_{m=2} = \{(34415.70-1.70)/\ 2\}\ /\ 0.57}$$
$$\mathbf{F_{m=2} = 30187.72}$$

Using the Microsoft Excel function for the F-distribution we determine the probability $\mathbf{P_{m=2}}$:

$$\mathbf{P_{m=2} = FDIST(30187.72, 2, 3)}$$
$$\mathbf{P_{m=2} = 3.50E\text{-}07}$$

Finally we calculate the level of confidence $\mathbf{1\text{-}\ P_{m=2}}$:

$$\mathbf{1\text{-}P_{m=1} = 1 - 3.50E\text{-}07}$$
$$\mathbf{1\text{-}P_{m=1} = 0.99999}$$

The confidence level is high (more than 99.99 %). We conclude that <u>*at least one*</u> of the parameters (b_2 or b_3) in the subset is non-zero. However, we <u>*can not be sure*</u> *that both quadratic and cubic terms are significant*.

3). Significance test of an individual coefficient in the regression model

Here the question to answer is: "With what confidence level can we state that the ith coefficient $\mathbf{b_i}$ in the model is significant?". The corresponding F-number is:

$$\mathbf{F_i = b_i^2 / [se(b_i)]^2} \qquad \textbf{(24)}$$

$\mathbf{se(b_i)}$ is the *standard error* in the individual coefficient $\mathbf{b_i}$ and is part of the *ANOVA output* (see Table 4a). The corresponding probability

$$\mathbf{P_i = FDIST(F_i, 1, n - k)} \qquad \textbf{(25)}$$

leads us to the confidence level $\mathbf{1\text{-}\ P_i}$ at which we can state that ***coefficient*** $\boldsymbol{b_i}$ ***is significant***. If this level is lower than desired one we say that ***coefficient*** $\boldsymbol{b_i}$ ***is insignificant.*** $\mathbf{F_i}$ ***is <u>not</u> part of spreadsheet regression output***, but might be calculated by hand if needed.

However, there is another statistics for testing individual parameters, which is ***part of ANOVA*** (see Table 4a):

$$\mathbf{t_i = b_i / se(b_i)} \qquad \textbf{(26)}$$

The $\mathbf{t_i}$ - number is the square root of $\mathbf{F_i}$ (expression (24)). It has a **Student's distribution** (see [1, p. 381] for the analytical form of the distribution). *The corresponding probability is numerically the same as that given by (25)*. There is a statistical function in Microsoft Excel which allows one to determine $\mathbf{P_i}$ (***part of ANOVA*** (see Table 4a)):

$$\mathbf{P_i = TDIST(t_i, n\text{-}k, 2)} \qquad \textbf{(27)}$$

Parameters of the function (27) are: the number of degrees of freedom df ($df_E = n - k$) and *form* of test (TL=2). If TL=1 a result for a *one-tailed* distribution is returned; if TL=2 two-tailed distribution result is returned. An interested reader can find more information about the issue in ref. [1]

Example

In our illustration $\mathbf{P_0=0.7881}$ and $\mathbf{P_3 = 0.6731}$ (see Table 4a) corresponds to fairly low confidence levels, $\mathbf{1 - P_0 = 0.2119}$ and $\mathbf{1 - P_3 = 0.3269}$. This suggests that parameters $\mathbf{b_0}$ and $\mathbf{b_3}$ are not significant. The confidence levels for $\mathbf{b_1}$ and $\mathbf{b_2}$ are high ($\mathbf{1 - P_1 = 1 - 0.0282 = 0.9718}$ and $\mathbf{1- P_2 = 1 - 0.0001 = 0.9999}$), which means that they are significant.

In conclusion of this F-test discussion, it should be noted that in case we remove even one insignificant variable from the model, we need to test the model once again, since coefficients which were significant in certain cases **might** become insignificant after removal and visa versa. It is a good practice to use a reasonable combination of all three tests in order to achieve the most reliable conclusions.

Confidence interval

In the previous section we were obtaining confidence levels given F-numbers or t-numbers. We can go in an opposite direction: given a desired minimal confidence level **1-P** (e.g. 0.95) calculate the related F- or t-number. Microsoft Excel provides two statistical functions for that purpose:

$$\mathbf{F_{(1-P)}=FINV\ (P,\ df_q,\ df_v)} \qquad \mathbf{(28)}$$

$$\mathbf{t_{(1-P)}=TINV(P,\ df\ )} \qquad \mathbf{(29)}$$

$\mathbf{df_q, df_v}$ **-** degrees of freedom of numerator and denominator, respectively (see (15))
df **-** degree of freedom associated with a given t-test (varies from test to test)

NOTE: in expression (29) **P** is the probability associated with so called ***"two-tailed"*** Student's distribution. A ***"one- tailed"*** distribution has the different probability α. The relationship between the two is:

$$\mathbf{\alpha=P/2} \qquad \mathbf{(30)}$$

Values of **F-numbers** and **t-numbers** for various probabilities and degrees of freedom are tabulated and can be found in any text on statistics [1,2,3,4] . Usually the "one-tailed" Student's distribution is presented.

Knowing the **t-number** for a coefficient $\mathbf{b_i}$ we can calculate the numerical interval which contains the coefficient $\mathbf{b_i}$ with the desired probability $\mathbf{1\text{-}P_i}$:

$$\mathbf{b_{L,\,(1-Pi)} = b_i - se(b_i) * t_{(1-Pi)}} \quad \text{(lower limit)} \qquad \mathbf{(31)}$$

$$\mathbf{b_{U,\,(1-Pi)} = b_i + se(b_i) * t_{(1-Pi)}} \quad \text{(upper limit)} \qquad \mathbf{(31a)}$$

$$\mathbf{t_{(1-Pi)} = TINV(P_i,\, n\text{-}k)} \qquad \mathbf{(32)}$$

The standard errors for individual parameters $\mathbf{se(b_i)}$ are part of ANOVA (Table 4a). The interval $\mathbf{[b_{L,\,(1-Pi)};\ b_{U,\,(1-Pi)}]}$ is called the **confidence interval** of parameter $\mathbf{b_i}$ with the $\mathbf{1\text{-}P_i}$ *confidence level*. The upper and lower limits of this interval at a *95% confidence* are listed in the ANOVA output by default (Table 4a; columns *"Lower 95%"* and *"Upper 95%"*). If in addition to this default, the confidence interval at a confidence other than 95% is desired, the box *"Confidence level"* should be checked and the value of the alternative confidence entered in the corresponding window of the *Regression input dialog box* (see Fig. 1).

<u>Example</u>

For the unrestricted model (1b), the lower and upper 95% limits for intercept are "-5.1413" and "6.1872" respectively (see Table 4a). The fact that with the 95% probability zero falls in this interval is consistent with our conclusion of insignificance of $\mathbf{b_0}$ made in the course *of F-testing of individual parameters* (see *Example* at the end of previous section). The confidence intervals at the 95% level for $\mathbf{b_1}$ and $\mathbf{b_2}$ do not include zero. This also agrees with the F-test of individual parameters.

In fact, analysis whether zero falls in a confidence interval could be viewed as a different way to perform the F-test (t-test) of individual parameters and must not be used as an additional proof of conclusions made in such a test.

<u>Regression statistics output</u>

The information contained in the *"Regression statistics"* output characterizes the "goodness" of the model as a whole. Note that quantities listed in this output can be expressed in terms of the regression F-number $\mathbf{F_R}$ (Table 4) which we have already used for the significance test of *all coefficients*.

Example

For our unrestricted model (1b) the output is:

Table 7. Regression statistics output*

Multiple R	0.99999882
R Square (R^2)	0.99999765
Adjusted R Square (R^2_{adj})	0.9999953
Standard Error (S_y)	0.75291216
Observations (n)	7

** - Corresponding notation used in this handout is given in parenthesis*

Standard error (S_y):

$$S_y = (MS_E)^{0.5} \qquad (33)$$

MS_E is an error variance discussed before (see expression (11)). Quantity S_y is an estimate of the *standard error* (*deviation*) of experimental values of the dependent variable **y*** with respect to those predicted by the regression model. It is used in statistics for different purposes. One of the applications we saw in the discussion of *"Residual output"* (Standardized residuals; see expression (2a)).

Coefficient of determination R^2 (or R Square):

$$R^2 = SS_R / SS_T = 1 - SS_E/SS_T \qquad (34)$$

SS_R, SS_E and SS_T are *regression, residual (error)* and *total* sum of squares defined by (7), (6a) and (3a) respectively. Th*e coefficient of determination* is a measure of the regression model as whole. The closer R^2 is to one, the better the model (1) describes the data. In the case of a perfect fit $R^2=1$.

Adjusted coefficient of determination R^2 (or Adjusted R Square):

$$R^2_{adj} = 1- \{SS_E / (n-k)\} / \{SS_T / (n-1)\} \qquad (35)$$

SS_E and SS_T are the *residual (error)* and the *total sum of squares* (see expressions (6a) and (3a)). The significance of R^2_{adj} is basically the same as that of R^2 (the closer to one the better). Strictly speaking R^2_{adj} should be used as an indicator of an adequacy of the model, since it takes in to account not only deviations, but also numbers of degrees of freedom.

Multiple correlation coefficient R:

$$\mathbf{R = (SS_R / SS_T)^{0.5}} \qquad \mathbf{(36)}$$

This quantity is just the square root of *coefficient of determination*.

Example

The fact that $\mathbf{R^2_{adj}}$ = **0.9999953** in our illustration is fairly close to 1 (see Table 7) suggests that overall model (1b) is adequate to fit the experimental data presented in Table 1. However, it does not mean that there are no insignificant parameters in it.

REGRESSION OUTPUT FORMULA MAP

For references, the following tables present a summary of the formula numbers for individual items in the Microsoft Excel Regression Output. Variables in parenthesis, introduced and used in this handout, do not appear in the output.

Table 8. Formula map of Regression statistics output

Multiple R	(36)
R Square (R^2)	(34)
Adjusted R Square (R^2_{adj})	(35)
Standard Error (S_y)	(33)
Observations (n)	

Table 9. Formula map of Residual output

Observation (j)	Predicted Y (y_j)	Residuals (r_j)	Standard Residuals(r'_j)
1	(1)	(2)	(2a)
2	(1)	(2)	(2a)

Table 10. Formula map of ANOVA output (part I)

	df	SS	MS	F	Significance F
Regression	(df_R) (14)	(SS_R) (7)	(MS_R) (13)	(F_R) (17)	(P_R) (18)
Residual (error)	(df_E) (12)	(SS_E) (6a)	(MS_E) (11)		
Total	(df_T) (10)	(SS_T) (3a)	(MS_T)* (9)		

*****- *not reported in Microsoft Excel* Regression output

Table 10a. Formula map of ANOVA output (part II)

	Coefficients (b_i)	**Standard Error ($se(b_i)$)**	**t Stat (t_i)**	**P-value (P_i)**	**Lower 95% ($b_{L,(1-Pi)}$)**	**Upper 95% ($b_{U,(1-Pi)}$)**
Intercept (b_0)			(26)	(25) (27)	(31)	(31a)
X Variable 1 (x_1)			(26)	(25) (27)	(31)	(31a)
X Variable 2 (x_2)…			(26)	(25) (27)	(31)	(31a)

LITERATURE

1. Afifi A.A., Azen S.P. “Statistical analysis. Computer oriented approach”, Academic press, New York (1979)
2. Natrella M.G. “Experimental Statistics”, National Bureau of Standards, Washington DC (1963)
3. Neter J., Wasserman W. “Applied linear statistical models”, R.D. Irwin Inc., Homewood, Illinois (1974)
4. Gunst R.F., Mason R.L. “Regression analysis and its application”, Marcel Dekker Inc., NewYork (1980)
5. Shoemaker D.P., Garland C.W., Nibler J.W. “Experiments in Physical Chemistry”, The McGraw-Hill Companies Inc. (1996)