BIOSTATISTICS
TOPIC 9: ANALYSIS OF CORRELATIONS
II. MULTIPLE REGRESSION ANALYSIS

PREDICTION AS A SCIENCE ?

Some years ago, I read a book which has a the following conversation (not precise words): "A bar man asks Andy Capp which one he would choose - money, power, happiness, or the ability to foretell the future? 'Foretell the future', Andy answers: 'that way I can make money. Money will bring me power, and then I will be happy' ".

It is probably fair to say that the dream of being able to predict the future is as old as human nature. Many of us normally disdain the notion of "fortune-telling", which is quite curious because science itself revolves around methodologies for telling the future. In fact, we merely use a different vocabulary. In contrast to fortune-telling, we talk about calculations instead of predictions, laws instead of fate, and statistical fluctuations instead of accidents. Yet, the aim of the scientific method is the same. From the observation of past events, we derive laws that, when verified, enable us to predict future outcomes.

Taking for instance, the concept that all animals die at the same age sounds implausible, if we measure age in years and months and days, but it becomes rather logical if we count the number of heartbeats. It is the only heartbeat that differs from animal to animal. Small ones, like mice, live about 3 years but their heartbeat is very rapid. Middle-size ones such as rabbits, dogs, sheep etc. have a slower heartbeat and live between 12 and 20 years. Elephants live more than 50 years but have a slow heartbeat. It is not surprised that a famous professor has claimed that "most mammals living free in nature (not in homes or zoos) have accumulated about one billion heartbeats on average when they die". It looks like we can predict life expectancy of animals from their heartbeats. But still, we need to have a systematic way of doing this. Modern science has given us the regression analytical technique to achieve this aim. We are going to discuss some practical aspects of this technique in this topic.

I. INTRODUCTION

In the last topic we consider regression model with one independent variable. Quite frequently in the analysis of data, we wish to study the dependence of a random variable Y on several variables $x_1, x_2, \ldots, x_p$. In this topic, we will extend the idea to include more than one independent variable in the equation. The technique is called **multiple linear regression**.

For the purpose of illustration, let us now consider a numerical data set resulting from a study which examined the heat generated during the hardening of Portland cements, which was assumed to be a function of the chemical composition, the following variables were measured:

x_1 : amount of tricalcium aluminate
x_2 : amount of tricalcium silicate
x_3 : amount of tetracalcium alumino ferrite
x_4 : amount of dicalcium silicate
Y : heat evolved in calories per gram of cement.

Table 1: Observed data

i	x_1	x_2	x_3	x_4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3

13	10	68	8	12	109.4

Variables x_1, x_2, x_3 and x_4 were measured as percent of weight of the clinkers from which the cement was made. Under the assumption that the heat generated during hardening is a linear function of the four variables, we postulate the model:

$$\begin{aligned}\mu_Y &= \mu_Y(x_1, x_2, x_3, x_4) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4\end{aligned}$$

This model postulates that at the point $(x_{1i}, x_{2i}, x_{3i}, x_{4i})$ (ith row of the data matrix), the expected value (or mean) of the heat is equal to $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$. The measured value y_i is thus considered as a realisation of a random variable Y_i, which consists of the above mean plus a random deviation e_i:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + e_i$$

The random error e_i are usually assumed to be mutually independent and to follow a normal distribution with mean 0 and a common variance σ^2.

In general, the coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ and β_4 are unknown. Regression methods are used to estimate them and to test hypotheses about them. The **intercept** β_0 is the value of μ_Y at the point (0, 0, 0, 0). β_1, β_2, β_3 and β_4 are called **partial regression coefficients.** They can be interpreted as follows: μ_Y increases by β_j if x_j increases by 1 while all other x-variables remain unchanged. Often, β_0 is referred to as the constant, which comes from the fact that β_0 can be considered as the partial regression coefficient for a variable x_0 that takes always the constant value $x_{0i} = 1$. We prefer to call β_0 the intercept.

Some more terminology: Y is called **dependent variable** or **response variable**. The x_j are called **regressors** or **independent variables** (although they need not be independent from each other in a statistical sense). It is important to remark that x_j **are not considered as random variables** but rather as variables whose values have been fixed by the investigator. If the x_j are random, then the conclusions are conditional on the realised values.

In practice, the linear model will hardly ever be valid exactly. In many cases, however, in the domain considered, it is good approximation to the real, more complex world. Moreover, it is rarely be possible to know, or take into account, all the quantities influencing the mean μ_Y. The deviation e_i from the linear model can be thought of as the sum of many unknown or uncontrollable influences and, possibly, of a measurement error.

Let us now summarise the model as follows:

(a) The mean μ_Y of the random variable Y depends linearly on the regressors x_1, x_2, x_3 and x_4 :

$$\begin{aligned}\mu_Y &= \mu_Y(x_1, x_2, x_3, x_4) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4\end{aligned}$$

(b) In each point $(x_{1i}, x_{2i}, x_{3i}, x_{4i})$, the deviations e_i from μ_Y are normally distributed with mean 0 and constant variance σ^2:

$$e_i \sim N(0, \sigma^2)$$

(c) The deviations e_i are mutually independent.

II. LEAST SQUARES ESTIMATION AND RESIDUALS

2.1. TYPICAL OUTPUT FROM A COMPUTER PROGRAM

The procedure of estimation of parameters in multiple regression analysis is complex. Nowadays, this mechanical task is normally handled by a computer program. The following output was produced by the SAS statistical analysis system:

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	2667.89944	666.97486	111.479	0.0001
Error	8	47.86364	5.98295		
C Total	12	2715.76308			

Root MSE	2.44601	R-square	0.9824
Dep Mean	95.42308	Adj R-sq	0.9736
C.V.	2.56333		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	62.405369	70.07095921	0.891	0.3991
X1	1	1.551103	0.74476987	2.083	0.0708
X2	1	0.510168	0.72378800	0.705	0.5009
X3	1	0.101909	0.75470905	0.135	0.8959
X4	1	-0.144061	0.70905206	-0.203	0.8441

In the following sections, we will discuss the significance as well as meaning of these numerical values.

2.2. LEAST SQUARES ESTIMATION

The unknown coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ and β_4 are estimated as follows: for each observation y_i, we form the deviation from the unknown mean μ_Y at the point $(x_{1i}, x_{2i}, x_{3i}, x_{4i})$. The sum of the *n* squared differences:

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \beta_4 x_{4i})^2$$

will be called the **sum of squares**. The latter is considered as a function of the $p + 1 = 4 + 1 = 5$ parameters $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and, by principle of least squares, we determine that particular 4-dimensional hyperplane (i.e. the value of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$) for which *S* is minimal. The value b_0, b_1, b_2, b_3, b_4 which minimise S are called **least square estimates** for $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. Thus we have:

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{3i} - \beta_4 x_{4i})^2 \leq S$$

SSE is variously called the **error sum of squares** or **residual sum of squares** or **minimum sum of squares**. The computation of the coefficients b_j requires the solution of a system of linear equations in 5 variables. However, we are not going to discuss this technical aspect in this note.

Now, having obtained b_j, we can compute for observation the estimated value (often called **predicted** value or **fitted** value):

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + b_4 x_{4i}$$

which indicates the value of the estimated hyperplane at the point $(x_{1i}, x_{2i}, x_{3i}, x_{4i})$.

2.3. RESIDUAL ANALYSIS

We will also compute the residuals, which is the deviations of the measured values y_i from their predicted counterparts:

$$\hat{e}_i = y_i - \hat{y}_i$$

SSE can then be written as:

$$SSE = \hat{e}_i^2 = (y_i - \hat{y}_i)^2$$

If n is large, we can consider the list of residuals $\hat{e}_i$ (i = 1, 2, 3, . . ., n) assuming the validity of the model, approximately as a random sample from a normal distribution with mean 0. The **variance of the residuals** is estimated by:

$$s^2 = \frac{SSE}{n - p - 1}$$

i.e **the standard deviation of the residuals** is:

$$s = \sqrt{\frac{SSE}{n - p - 1}}$$

where p is the number of x-variables (in our example, p = 4).

s is also called the **standard error of regression**.

In our example, we obtained the equation:

$$\hat{\mu}_Y = 62.418 + 1.551x_1 + 0.510x_2 + 0.102x_3 - 0.144x_4$$

Based on this equation, we can calculate the predicted value of y (denoted by $\hat{y}$) for any given values of x_1, x_2, x_3 and x_4. This predicted value together with observed value is tabulated in the Table 2.

Table 2: Observed, predicted values of y and residuals.

i	y_i	$\hat{y}_i$	$\hat{e}_i$
1	78.5	78.495	0.005
2	74.3	72.789	1.511
3	104.3	105.971	-1.671
4	87.6	89.328	-1.728
5	95.9	95.648	0.252
6	109.2	105.274	3.926
7	102.7	104.148	-1.448
8	72.5	75.676	-3.176
9	93.1	91.722	1.378
10	115.9	115.619	0.281
11	83.8	81.810	1.990
12	113.3	112.326	0.974
13	109.4	111.693	-2.293

with a residual sum of squares of :

$$SSE = \hat{e}_i^2 = \left(y_i - \hat{y}_i\right)^2 = 47.945$$

and the standard deviation of residuals:

$$s = 2.45.$$

The method of least squares gives a "best-fitting" hyperplane whether the assumptions stated earlier is satisfied or not. However, if the assumptions are valid, the least squares estimators have some important properties: the quantities b_0, b_1, b_2, b_3, b_4 are unbiased estimates of the parameters $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. Among all possible unbiased estimators, which are linear functions of the y_i, the least squares estimators have the smallest variance. The quantity s^2 is an unbiased estimator of σ^2. Moreover, from the assumption that the e_i

are independent and normally distributed, it follows that the least squares estimators b_0, b_1, b_2, b_3, b_4 are also normally distributed.

In order to validate the assumptions made, it is very important to examine the residuals. A histogram of the residuals is a possible aid for visualising possible violations of the normality assumptions. Also important is a scatterplot of the pairs $(\hat{y}_i, \hat{e}_i)$, i.e. of the residuals versus the predicted values, which often reveals whether or not the underlying linear model is correct.

If the assumptions are satisfied, the residuals should scatter around 0 and there should be no dependency on the predicted value of y. Non-constant variance of the residuals, non-linear dependency and other forms of violations of the model assumptions can be uncovered by means of this simple plot.

Table 2 also gives a list of the predicted values of y and residuals computed from the actual equation relating y on the 4 x-variables. .A plot of residuals and predicted values of y $(\hat{y}_i, \hat{e}_i)$ is shown in Figure 1. It seems that there is no relationship between the two variables, and hence the assumption of random residuals seem to satisfied.

Occasionally, scatterplot of the residuals versus individual independent x variables can also furnish information about non-linear dependencies. If a residual is extremely large in absolute value, we may be suspicious of an outlier, or the linear model may not be valid for that particular point. We are not going to give more details on residual analysis since it is an extensive topic of statistical research.

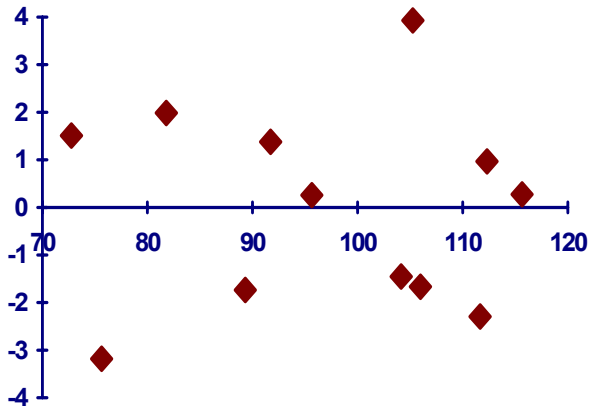

Figure 1: Plot of residuals versus predicted value of *Y*.

III. ANALYSIS OF VARIANCE

An analysis of variance partitions the overall variation between the observations Y into variation which has been accounted for by the regression on X and residual or unexplained variation. Thus, we can say:

Total variation about the mean	=	Variation explained by regression model	+	Residual variation

In ANOVA notation, we can write equivalently:

$$\text{SSTO} = \text{SSR} + \text{SSE}$$

or,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Now, SSTO is associated with n-1 df. For SSR, there are five parameters (b_0, b_1, b_2, b_3 and b_4) in the model, but the constraint $\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = 0$ takes away 1df, hence it has finally 4 df. For SSE, there are n residuals (e_i); however, 5 df are lost because of two constraints on the e_i's associated with estimating the parameters β_0, β_1 , β_2 , β_3 , and β_4 by the normal equations see Topic 8).

We can assemble these data in an ANOVA table as follows:

Source	df	SS	MS
Regression	4	SSR = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MSR = SSR/4
Residual error	n - 5	SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MSE = SSE / (n-5)
Total	n - 1	SSTO = $\sum_{i=1}^{n}(y_i - \bar{y})^2$	

In our example, the corresponding numerical values for the analysis are:

Table 3: Analysis of variance of model

$$\hat{\mu}_Y = 62.418 + 1.551x_1 + 0.510x_2 + 0.102x_3 - 0.144x_4$$

Source	df	SS	MS
Regression	4	2667.90	666.97
Residual error	8	48.76	5.98
Total	12	2175.76	

In this example, the total sum of squares about the mean is $\sum_{i=1}^{13}(\hat{y}_i - \bar{y})^2 = 2715.76$ and the sum of squares "explained" by the regression equation is $\sum_{i=1}^{13}(\hat{y}_i - \bar{y})^2 = 2667.90$, which leaves a residual sum of squares of $\sum_{i=1}^{13}(y_i - \hat{y})^2 = 47.86$. Consequently, the proportion of total variation of *y* attributable to the model is 2667.90 / 2715.76 = 0.9824 or 98.24%. We will discuss the "adjusted R^2" in a later part of this topic.

Since the data set has 13 observations, therefore there are 12 df associated with the total sum of squares. On the other hand, the model has 5 parameters (including the intercept), therefore the number of degrees of freedom (df) associated with the equation is 4, hence the mean square due to the model is 2667.90 / 4 = 666.97. That makes the number of df associated with the residual sum of squares of 8 and hence its mean square is 47.86 / 8 = 5.98. By definition, the F test of significance of the model is given by F = 666.97 / 5.98 = 111.479 which is statistically highly significant, compared with the theoretical F value of 3.84 at $\alpha = 0.05$ with numerator and denominator df of 4 and 8, respectively.

The residual mean square (MSE) could be treated as an estimate of variance of the model. Thus, the "root MSE", which is actually the square root of MSE, can be treated as an estimate of standard deviation of the model e.g. $\sqrt{5.98} = 2.45$. Now, the observed

variance of y is $\sum_{i=1}^{13}(\hat{y}_i - \bar{y})^2$ /12 = 2715.76 / 12 = 226.31. After accounted for by the model (with 4 variables), the variance is only 5.98, a reduction of 97.3%. This is a respectable number !

3.1. OVERALL TEST OF SIGNIFICANCE

The linear model with p partial regression coefficients is only meaningful if at least one coefficient is different from zero (significant). We compare the full model

$$\mu_Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

with the reduced model

$$\mu_Y = \beta_0$$

by eliminating all p variables simultaneously, i.e. by putting:

$$\beta_0 = \beta_1 = \ldots = \beta_p = 0$$

The least squares estimate of β_0 turns out to be the sample mean of Y, i.e. $b_0 = \overline{Y}$; the corresponding SSE is:

$$SSE_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

This is in fact referred to as "Total" SS in the ANOVA table.

The question is then "do the four variables significantly explained the observed variation of the data?". This question can be addressed by calculating the overall F statistic which is: F = 666.97/5.98 = 111.29. This value is compared (at $\alpha = 0.05$) with 95% quantile of the F distribution with 4 and 8 DF, respectively. Now, $F_{0.95,4,8} = 3.84$. Thus, there is strong evidence that at least one of the partial regression coefficients is different from zero, and we can not dispense with all regressors.

All major computer programs for regression analysis calculate the ANOVA table for the overall test of significance and the associated F statistic. In contrast to our terminology, the SS in the ANOVA table are virtually labelled as follows in the computer printout: "TOTAL" (for the reduced model $\mu_Y = \beta_0$), "RESIDUAL" (for the full model $\mu_Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$) and "REGRESSION" (for the sum of squares associated with the

reduction). Our terminology has the advantage of reminding the user that actually mathematical *models* are being compared and not just the sums of squares.

3.2. COEFFICIENT OF DETERMINATION

In addition to the overall F value, one often computes the coefficient of determination R^2, which is defined as follows:

$$R^2 = \frac{SSE_0 - SSE_4}{SSE_0} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

(note: the subscript *k* in *SSE* refers to the SSE due to *k* variables in the model; thus, SSE_4 referred to the SSE due to 4 variables in the model, while SSE_0 refers to the SSE of the model without any independent variable)

This quantity is often interpreted as the proportion of the variability of Y explained by the regression on $x_1, x_2, \ldots, x_p$. In our examples, $R^2 = 0.982$, which says that 98.2% of the variability (not variance) of Y is accounted for the independent variables x_1 to x_4. If no linear dependency exists, then R^2 lies near 0; in this case of a strong linear dependency, however, near 1. When interpreting the coefficient of determination, however, caution is advised, because for $n \leq p+1$, one has $\hat{y}_i = y_i$, and thus $R^2 = 1$. Interpreting R^2 is meaningful only when *n* is considerably larger than *p*.

The root $R = \sqrt{R^2}$ is called the **multiple correlation coefficient** between Y and $x_1, x_2, \ldots, x_p$. R is equal to the correlation coefficient of the pair $(\hat{y}_i, y_i)$. In our example, R = 0.991.

3.3. ADJUSTED R^2

Some statisticians have suggested that the coefficient of determination should be modified to recognise the number of independent variables in the model. The reason is that this coefficient can generally be made larger if additional independent variables are added

to the model. To see this, note that SSE tends to be smaller with each additional independent variable, while SSTO remains fixed. A measure that recognises the number of independent variables in the model is called the adjusted coefficient of determination and is denoted by R_a^2:

$$R_a^2 = 1 - \frac{n-1}{n-p} \times \left(\frac{SSE}{SSTO} \right)$$

In our example, the adjusted R_a^2 is 0.974, which is not greatly different to that of 0.982.

IV. MODEL BUILDING AND ANALYSIS OF VARIANCE

Once the multiple regression equation has obtained, we would like to know whether it is necessary to include all independent variables in the equation. We can single out two cases in which a regressor can be removed from the model without any loss of information:

(a) There is no relationship between x_j and Y;

(b) The influence of x_j on Y is affected through other variables. This possibility can be illustrated most simply by means of a simulated example. Assume that the following model is valid:

$$y = x_1 + 2x_2$$

Assume, at the same time, that the variable $x_3 = x_1 + x_2$ has been measured too. We can therefore also write the model as:

$$y = x_2 + x_3$$

or as $y = 2x_3 - x_1$

finally, we can also describe y as a linear function of all three regressors, e.g.

$$y = 2x_1 + 3x_2 - x_3$$

In this model, one variable is clearly redundant - even though y is functionally dependent on x_1, x_2 and x_3. In practice it is often difficult to recognise such functional dependencies between the regressors. Often a functional relationship is confounded with a measurement error, or it is not exactly linear.

In both cases, we speak of redundancy: the variable x_j in case (1) or one of the variables x_1, x_2, x_3 in case (2) can be removed from the model without loss of information, i.e. it is redundant. We try to simplify the model through elimination of redundant regressors. As possibility 2 shows, however, redundancy does not mean that the regressor has no influence.

The removal of variable x_j from the model is done by setting the parameter β_j equal to zero. For simplicity of notation we assume that the variables are ordered in such a way that the parameters to be set equal to zero are the last ones. The model with all p regressors

$$\mu_Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_p x_p$$

will be called the full (or alternative model). By the so-called linear restriction, i.e. by setting, say:

$$\beta_3 = \ldots = \beta_p = 0$$

a simplified model

$$\mu_Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

is obtained. The latter is called **reduced model** or **null model**.

The full and reduced models can be compared as follows. For both models one estimates the parameters and computes the corresponding residual sum of squares. The latter shall be denoted by SSE_p (full model) and SSE_r (reduced model), respectively. Since the adjustment of the plane becomes worse by the elimination, one always has $SSE_p \leq SSE_r$. We now form the ratio:

$$F = \frac{(SSE_r - SSE_p)/(p-r)}{SSE_p/(n-p-1)}$$

Under the hypothesis $\beta_3 = \ldots = \beta_p = 0$, this statistic is distributed according to the F distribution with (p - r) degrees of freedom in the numerator [df] and (n - p - 1) df in the denominator. If the computed F ratio is smaller than the (1 - α) quantile of the F distribution with (p - r) and (n - p - 1) df, then the null hypothesis is accepted. Otherwise, we retain the full model.

The residual sum of squares and the F test are usually summarised in an analysis of variance table as follows:

Model	Minimum SS	DF
Reduced model $\beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q$	SSE_q	n - q - 1
Full model $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$	SSE_p	n - p - 1
Reduction $\beta_{q+1} = \ldots = \beta_p = 0$	SSE_q - SSE_p	p - q

It should be noticed that the correctness of the *F*-test depends on the assumptions of the model. In many cases it is appropriate to view the F-value merely as a descriptive measure of the difference between the two models. This is particularly true if a series of model comparisons is carried out on one and the same set of data. We will return to this issue in a next section.

TEST OF PARTIAL HYPOTHESES

In the second special case $q = p - 1$, mentioned in the above section, the reduced model is:

$$\mu_Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

which is obtained from the full model by imposing the restriction $\beta_p = 0$. For simplicity we again assume that it is the redundancy of the *p*th variable that needs to be tested. As in the

overall hypothesis, we compute the two minimum SS $S_{\min}^{p}$ (full model) and $S_{\min}^{p-1}$ (reduced model). By means of the ratio:

$$F = F\left(\beta_p = 0\right) = \frac{SSE_{p-1} - SSE_p}{SSE_p / \left(n - p - 1\right)}$$

called the partial F-statistic. We test the redundancy of x_p in the regression model. Under the validity of the null hypothesis $\beta_p = 0$, the test statistic is distributed as F with 1df in the numerator and (n - p - 1) in the denominator.

By comparing the realised F value with the (1- α) quantile of the $F < F_{1-\alpha,(1,n-p-1)}$, we accept the simpler reduced model. It is, however, not possible to conclude from this that x_p has no influence on Y, since the influence x_p can may be represented by some other regressors. The only thing we can say is that the full set of regressors does not describe the linear relationship better than the reduced set.

If several partial F-values are computed, we recommended their use merely for descriptive purpose. In particular, it is not correct to conclude from the non-significance of several partial regression coefficients that these may simultaneously be removed from the model. Indeed, the elimination of one single regressor can strongly influence by the other coefficients.

Many computer programs give *partial F-statistics* on n - p - 1 df. rather than partial F-statistics for testing the redundancy of a single variable. The relationship between the two statistics is simple F = t^2, reflecting the fact that F with 1 df in the numerator and m df in the denominator is the same as the square of t with m df.

STANDARD ERROR OF REGRESSION COEFFICIENTS

From the partial F-value, the *standard error of* b_j can be estimated as follows:

$$SE(b_j) = \frac{|b_j|}{\sqrt{F(\beta_j = 0)}}$$

In our example, one can thus supplement the estimated regression equation with a list of partial F values and standard errors of the partial regression coefficients as shown in the following table.

Table 4: Estimated regression coefficients, standard error and partial F test.

j	b_j	SE(b_j)	F($\beta_j = 0$)
0	62.418	71.59	0.760
1	1.551	0.745	4.331
2	0.510	0.724	0.496
3	0.102	0.755	0.018
4	-0.144	0.710	0.041

The rather large SE (compared to the absolute values of the coefficients) and the small F values lead one to suspect that the set of variables can be reduced in some way. An approach to eliminating redundant variables will be discussed in the next section.

VI. SELECTION OF A SUBSET OF REGRESSORS

The selection of a suitable subset of regressors is often difficult. Different algorithms may yield different results. It can also happen that different subsets of equal size yield results of practically the same quality. In our example, we could eliminate one of the four independent variables and would obtain for the four reduced models the following coefficients of determination:

Table 5: Comparison of various models by using R^2

Model without	R^2
x_1	0.973
x_2	0.981
x_3	0.982
x_4	0.982

Since none of the models with three regressors results from one of the others through a linear restriction, they can not be compared to each other by an analysis of variance. According to the coefficient of determination, three of the four models containing three regressors would be about equally good.

Situations such as this are especially likely to occur when the regressors are strongly related to each other. The relations among the x_j are usually described by a correlation matrix, although the values taken by x_j are considered as fixed for the purpose of least square estimation.

Because of the noncompatibility of models with an equal number of regressors one often uses hierarchical algorithms for selecting a subset of independent variables. We describe here the method of *backward elimination*. Starting with the full model with p variables, we first eliminate the variable with the smallest partial F value. In the remaining variables, i.e. we compare the model with p-1 variables with all models which result from

elimination of one additional variable. Again, the variable with the smallest partial F value is eliminated, etc. Again, the variable with the smallest partial F value is eliminated, etc. until in the end all regressors have been eliminated from the model. Applied to our example, the backward elimination yields the following results (MPF = minimum partial F value).

Step 1: Full model

Variable	b_j	SE(b_j)	F($\beta_j = 0$)	
x_1	1.551	0.745	4.33	
x_2	0.510	0.724	0.50	
x_3	0.102	0.755	0.02	MPF
x_4	-0.144	0.710	0.04	
Intercept	62.418			

$R^2 = 0.982$, Std error of residual = 2.448

Step 2: Elimination x_3

Variable	b_j	SE(b_j)	F($\beta_j = 0$)	
x_1	1.452	0.117	154.01	
x_2	0.416	0.186	5.03	
x_4	-0.237	0.173	1.86	- MPF
Intercept	71.648			

$R^2 = 0.982$, Std error of residual = 2.309

Step 3: Elimination x_4

Variable	b_j	SE(b_j)	F($\beta_j = 0$)

x_1	1.468	0.121	146.52 - MPF
x_2	0.662	0.046	208.58
Intercept	52.577		

$R^2 = 0.979$, Std error of residual = 2.406

Step 4: Elimination x_1

Variable	b_j	SE(b_j)	F($\beta_j = 0$)
x_2	0.789	0.168	21.96
Intercept	57.424		

$R^2 = 0.666$, Std error of residual = 9.077

Eliminating x_2 in step 5 finally yields the estimate $b_0 = \bar{y} = 95.423$ and Std error of residual = Std deviation of Y = 15.044.

It is now up to the us to make the important decision of how many regressors and which ones, we want to include in the model. Most programs offer the possibility of indicating a critical value $F_{\min}$. The algorithm is stopped as soon as there is no longer a partial F value smaller than $F_{\min}$. In the hierarchical sequence of tests it is, however, not possible to consider the partial F values as independent. Therefore, the use of the partial F value only in the descriptive sense for the purpose of determining an ordering among the independent variables. A possible stopping criterion is an abrupt change in R^2. In our example, the coefficient of determination decreases abruptly after three steps, this sharp bend is an indication that the two regressors eliminated first are redundant. The remaining model is based on x_1 and x_2 only; the regression equation (in brackets standard errors of the b_j) is estimated as:

$$\hat{\mu}_Y = 52.577 + 1.468x_1 + 0.662x_2$$
$$(0.121)(0.046)$$

This model can now be compared with the full model. The ANOVA table as shown below give the value of the test statistic as:

$$F = \frac{9.96/2}{47.945/8} = 0.83$$

lies below the 0.95 quantile of the F distribution with 2 and 8 df [$F_{0.95,2,8} = 4.46$]. The simplification of model to 2 regressors appears to be justified.

Model	SSE	DF
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	57.904	10
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$	47.945	8
Reduction	9.960	2

It must be emphasized that the solution thus found need by no means be the only correct one. A model with (x_1, x_4) instead of (x_1, x_2) for example, yields practically the same coefficient of determination. It remains up to the user to decide, based his/her knowledge of the subject, which regressors he/she wants to use to describe the dependent variable.

Moreover, strictly speaking, the test given above is not entirely correct, since it violates the principle that **hypotheses should not be generated and tested on the same data**.

The following table gives a list of the observed and predicted value of the dependent variable (y_i and $\hat{y}_i$) , and the residuals $e_i = y_i - \hat{y}_i$ for the model with (x_1, x_2). The multiple correlation for this model is plotted by the pairs y_i and $\hat{y}_i$. This figure, at the same time, allows an examination of the residuals: the horizontal (or vertical) distance of (y_i and $\hat{y}_i$) to the straight-line $\hat{y}_i = y_i$ (slope = 1, angle = 45o) corresponds precisely to the residual $e_i = y_i - \hat{y}_i$. No violation of the assumptions can be detected.

Table 6: Predicted values and residuals for the

model $\mu_Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

i	y_i	$\hat{y}_i$	$\hat{e}_i$
1	78.5	80.074	-1.574
2	74.3	73.251	1.049
3	104.3	105.815	-1.515
4	87.6	89.258	-1.658
5	95.9	97.292	-1.393
6	109.2	105.152	4.048
7	102.7	104.002	-1.302
8	72.5	74.575	-2.075
9	93.1	91.275	1.825
10	115.9	114.538	1.362
11	83.8	80.536	3.264
12	113.3	112.437	0.863
13	109.4	112.293	-2.893

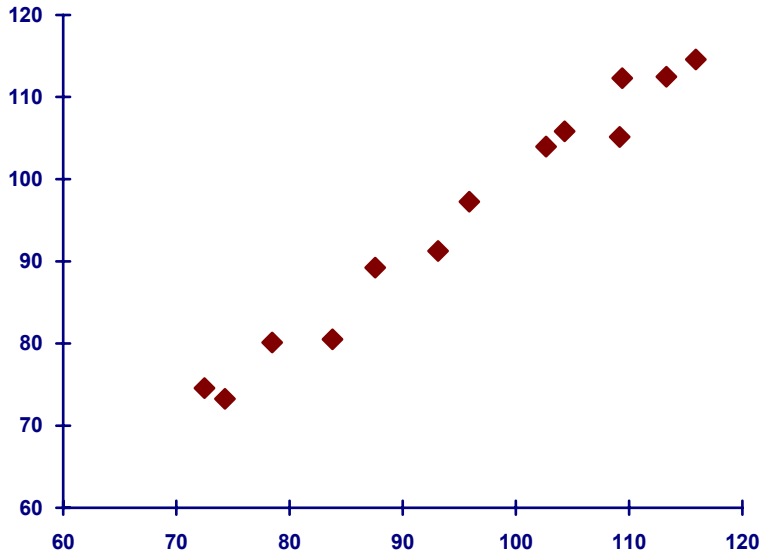

Figure 2: Multiple Correlation in the Model $\mu_Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

VII. COMMON QUESTIONS AND ANSWERS

Question: *What should I do if the assumptions of the model (normal distribution of the deviations) are not met*?
Answer: In practice, one often encounters problems in which the assumptions for statistical testing are not satisfied. Purely numerically, one can, of course still compute a "best fit" hyperplane. But the statistical tests are no longer correct. Since in this topic we consider multiple regression merely as a tool, we refer here to the relevant literature.

Question: *Apart from backward elimination, are there any other procedures for selecting subsets of variables*?
Answer: In statistical program libraries, the following additional algorithms are often used:

(a) Forward selection: In this method, one seeks in step 1 a first variable which, taken alone, yields the highest F value. In step 2, among all pairs of regressors which contain the already selected variable, one finds the one which maximises the overall F value. Analogously, in step 3 a third variable is added to the two already selected, and so on.

(b) Stepwise selection: This method is a mixture of the two already described. Step 1 and 2 are as in forward selection: subsequently however, prior to the selection of an additional variable, one always examines the partial F value of all variables already selected. If such a value falls below a certain limit (which is imposed by the user), the variable in question is again eliminated, whereupon another variable is newly included, and so on. The following table gives a list of all 15 possible subsets of regressors. On the basis of this table, we can trace all three algorithms.

Step No.	Backward	Forward	Stepwise
1	$x_1\, x_2\, x_3\, x_4$	x_4	x_4
2	$x_1\, x_2\, x_4$	$x_1\, x_4$	$x_1\, x_4$
3	$x_1\, x_2$	$x_1\, x_2\, x_4$	$x_1\, x_2\, x_4$
4	x_2	$x_1\, x_2\, x_3\, x_4$	$x_1\, x_2$
5			$x_1\, x_2\, x_4$

A comparison between backward elimination and forward selection shows the phenomenon, which is rather rare in practice, that the reversal of forward selection does not produce at all stages the same subset of regressors as backward elimination. The procedure stops at step 5, since in the next step x_4 would again be eliminated.

In the following, we will always use backward elimination, since at the beginning one performs the overall test. If this test turns out to be non-significant, then the decision is that the coefficients β_1 , $\beta_{2,\ldots,}\beta_p$ are all zero. In this case, any selection of subsets is unnecessary, whereas in the other procedures, from a number of purely redundant regressors a few can be selected, even though in reality there exists no influence.

Thanks to fast computers it is nowadays also possible to use so-called "**all subsets regression**" algorithms. If p regressors are available, an exhausive search implies that $\left(2^p-1\right)$ subsets have to be analysed - an almost astronomic task even if p is as small as 10. Of course, it is possible to restrict the number of subsets by heuristic considerations, but nevertheless, the amount of computation is very large.

Question: *In backward elimination, can one come up with an automatic stopping criterion that would avoid following the algorithm to the very end*?
Answer: In most computer programs, this is, indeed, possible. Backward elimination carried through to the end, however, may provide additional information about the importance of the individual regressors.

Question: *Are not the assumptions in our example is grossly violated? The variables* x_1 *to* x_4 *as one can see from the data matrix, care certainly not normally distributed*?
Answer: This is one of the most misunderstood issue among regression analysts or users. The assumption of normality is applied to the residual e_i, not to the x_i,. The latter is assumed to be fixed (not random) variable.

Question: *What about the numerical accuracy of the results? As is known from experience, the results of various regression programs do not always agree to all decimal digits*?
Answer: The numerical computation of the coefficients consists of mainly of the inversion of a matrix. In this connection, certain numerical problems may arise, especially when the regressors are strongly correlated among each other. It is therefore recommended to analyse

one and the same regression problem with different programs. If the results do not agree, a possible way out of this is to consider instead of the actual regressors, their principal components as independent variables. This option is provided in some advanced regression analysis programs.

Question: *Why should not we use the full model with all p variables included*? *Why should we make the fit worse by eliminating variables? Since* R^2 *always increases if we include more variables, it certainly can't hurt if we use all the predictors that are available?*
Answer: From a non-statistical viewpoint, this observation is right - inclusion of an additional variable will always improve the fit as measured by the coefficient of determination (R^2). The statistical question is whether or not the improvement is purely random. There are also other reasons for selecting a subset of regressors. For instance, we can argue that a simple model is preferable to a complicated model, provided that both models work equally well - a rule that is valid not only for statistical models. A very important reason is the instability of the parameter estimates if too many variables are included - instability in the sense of both poor numerical accuracy and large standard errors of the parameter estimates. This is clearly visible in the data set in our example: if we follow the backward elimination process, in step 1, all coefficients are highly unstable (in the sane of high standard errors), while after elimination of x_3 and x_4, the remaining coefficients are rather stable. This phenomenon occurs typically when the regressors are highly correlated among each other. Subset selection is, admittedly, only one possible way to handle this problem. Another interesting technique is "**ridge regression**", which trades the high variability of the parameter estimates for some (hopefully negligible) bias. This is not covered in our course, but interested readers can be referred to suitable text on the subject.

Question: *We have discussed the role played by different variables in a regression equation, and we have seen that some variables may be more important for predicting Y than others? Could not we do a similar thing for the observations? I mean, is it possible to measure to what extent the regression equation is determined by each observation?*
Answer: Methods for assessing the influence of individual observations on the regression have indeed been given much attention in recent years. In fact, we have reviewed these techniques briefly in Topic 8 (influential observation, Cook's statistic etc.). However, as mentioned in the section of residual analysis, a simple approach is, for instance, to omit observation number *i* and recalculate the regression equation based on the remaining *n*-1 data points. The influence of the *i*th observation can then be judged by the changes

occurring in the regression coefficients. For a nice introduction to these techniques, you are referred to a paper by Efron and Gong (1983).

Knowing the background of the data is often helpful to determine which observations need yo be looked at more closely for their influence. This is also true for the data in our example, in which the detailed knowledge of the experiment can be used to build an appropriate model and to select observations.

Question: *Should one rely on the coefficient of determination* $\left(R^2\right)$ *to assess the goodness-of-fit of the model?*

Answer: For interpreting R^2, it may be helpful to take notice the following theoretical result: under *no* conditions on the *x*-variables (i.e. their values may be random according to any law, or they may be fixed), and if the random variable Y does not depend on the *x*-variables, then under rather weak assumptions on the distribution of Y (especially symmetry and incorrelatedness of the realisations of y_i) the expected value of R^2 is $E\left(R^2\right)=\frac{p}{n-1}$. That is, for a sample size of $n = 21$ and $p = 10$ *x*-variables, we can expect to have $R^2 = 0.5$ by pure chance!!! So, as a simple rule of thumb, only the proportion of R^2 exceeding the expected value should be interpreted.

In connection with this result, it is also worth noting that the overall F test of significance (which is a functionally closely related to R^2) is valid under similar weak conditions. This means particularly that the overall F test does not require exact normality of the residuals, but broadly speaking, a symmetric distribution. For the partial F tests, it is not known to what extent their correctedness depends on strict normality. In general, however, these tests are more reliable the larger the sample size *n*.

Question: *If we have only one regression x, but the regression function was a quadratic function of x. Can this problem be approached by linear regression technique ?*

Answer: To answer this question, we must firstly clarify our terminology. In the linear regression model, $\mu_Y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$, the word "linear" refers actually to the fact that μ_Y is a linear function of the parameters β_j. The model that the question is about has the form: $\mu_Y = \beta_0 + \beta_1 x_1 + \beta x_1^2$, which can be handled within the framework of linear regression with two regressors, namely, x_1 and $x_2 = x_1^2$. On the other hand, the model $\mu_Y = \beta_0 + \beta_1 x_1 + \beta^2 x_2$m is not linear in the parameter β, and the linear regression technique can be applied.

VIII. SOME COMMENTS

Multiple regression analysis is one of the most often use (and abuse) statistical techniques, particularly in the medical literature. Some people seem to have disregard for scientific principle and manipulate the data to suit the hypothesis they generated. The following comments are extracted from D. Altman's book:

"It is not possible to discuss in detail many of the important issues that affect multiple regression analysis and its interpretation, but the following comments indicate areas of interest of difficulties:

When there is a large number of potential explanatory variables we expect some of them to be significant just by chance. There is no completely satisfactory way of searching for the most suitable model without incurring the penalty of an over-optimistic answer. With many candidates for inclusion in the model, some researchers use the results of univariate analyses to decide which variables should be explored in the multivariate analyses. This strategy saves nothing with forward stepwise regression, but may dramatically cut computing time (and costs) for backward stepwise or all subsets regression. I do not recommend pre-selection, but if it is used, selection should be based on a lax criterion, say $p < 0.2$ or even higher, because variables that may contribute to a multiple regression in unforeseen ways due to complex interrelationships among the variables. As an example, the cyclic fibrosis data set gave $p = 0.27$ for BMP on its own, but $p = 0.019$ for the same variable in the multiple regression model.

Because of the multiple testing at each step, a model derived by stepwise regression is likely to be over-optimistic with respect to the importance of each variable and the goodness-of-fit, particularly in small samples. When the number of explanatory variables is large and the number of observation is small, it is possible to find a model that appears to fit remarkably well. However, a model containing, say 7 variables, fitted to 18 observations will be extremely unreliable. One solution is to suggest that multiple regression should not be applied to small data sets. In addition, it should be decided in advance the maximum size of the model that is acceptable. I have found that the square root of the sample size a useful rule of thumb here, but even that may be over-generous. Alternatively, it is sometimes suggested that the number of variables examined should be restricted. Again, there is no

rule, but a guideline might be to look at no more than *n*/10 variables, where *n* is the sample size.

When the sample size is very large, statistical significance can be achieved for a tiny effects. For example, Rantakallio and Makinen (1984) fitted a model to data from 9795 infants on the number of teeth at one year of age. Six of the 15 variables were statistically significant ($p < 0.05$), one being the sex of the child ($p<0.001$). The regression coefficient was -0.051, indicating a mean difference of one-twentieth of a tooth in favour of boys. The value of R^2 for this model was only 3.1%.

Automatic procedure for selecting a model are useful, but a degree of common sense is required. For example, sometimes there is an accumulative evidence that a particular variable is prognostically important for the outcome being analysed. It is not possible to omit, say, age or smoking in such circumstances because p was "only" 0.07.

The question of how well the model fits the data was discussed. The R^2 and adjusted R^2 are one way of assessing goodness-of-fit, but they are measures of the correlation between observed and predicted values of Y (the dependent variable). We can not get any idea of the accuracy of prediction for an individual from the significant variables nor from R^2, however large it is. As with ordinary linear regression, the residual standard deviation gives a measure of the discrepancies between the observed and predicted Y values, from which a 95% prediction or confidence limit can be obtained.

Lastly, because of the risk that the model may be over-optimistic, it is desirable to assess the predictive capability of a model on a new independent set of data, but this is not usually possible."

IX. EXERCISES

1. The following data are the measurements of the height and weight of 10 men:

Height:	63	71	72	68	75	66	68	76	71	70
Weight:	145	158	156	148	163	155	153	158	150	154

(a) Find the linear regression of height on weight
(b) Find the linear regression of weight on height.
(c) Explain why the two equations (in (a) and (b)) are different?

2. Results of fitting a regression model $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$, based on $n = 32$ observations, are as follows:

$b_0 = -1.707$
$b_1 = 0.01681 \qquad SE(b_1) = 0.0003937$
$b_2 = -3.507 \qquad SE(b_2) = 0.4132$
SSR = 6444.8
SSE = 55.14.

(a) Interpret the coefficient b_1.
(b) Set up the ANOVA table. Calculate R^2 and interpret it.
(c) Obtain an estimate for σ^2.

3. In a regression study involving 3 independent variables, after fitting various models, the investigator summarised the data as follows:

Model	Regression Sum of Squares
x_1	2970.6
x_2	3654.8
x_3	3584.5

x_1, x_2	5123.8
x_1, x_3	5409.6
x_2, x_3	3741.3
x_1, x_2, x_3	5409.9
Total SS of observed Y	5426.2

Based on these results, which model would you adopt as the "final" or optimum model?

4. The following data are about the number of defective pipes in shipment (Y) and total number of pipes in a shipment (X) for 12 shipments.

X:	5	10	4	10	7	8	8	5	10	5	12	6
Y:	30	51	26	52	40	43	45	31	52	30	59	36

Fit a quadratic equation and interpret the data.

5. The following tabulation gives the region, number of beds (X_1) and number of admissions (Y) last year for each of 24 small acute-care hospitals:

Region A	X_1	Y	Region B	X_1	Y	Region C	X_1	Y
1	19	120	1	96	2958	1	76	2648
2	120	3374	2	48	1487	2	75	2757
3	49	2244	3	148	4700	3	84	2881
4	100	3606	4	101	3308	4	13	402
5	33	950	5	66	2696	5	40	1600
6	22	703	6	138	4845	6	69	1646
			7	25	1159	7	125	4825
			8	193	5692	8	13	370
			9	44	1576	9	32	987

(a) Use X_2 and X_3 to define the region as follows:

Region	X_2	X_3
A	1	0
B	0	1
C	0	0

Obtain the estimated regression function. Interpret the meaning of b_1, b_2 and b_3. Give an estimate of the mean number of admissions for 100 bed hospital in regions B and C. Does the mean admissions differ among the three regions for hospitals with a given number of beds? Comment.

(b) Set up an ANOVA table. Calculate $\sqrt{MSE}$. What does this number measure in this study.

6. For lung transplantation it is desirable for the donor's lungs to be of a similar size as those of the recipient. Total lung capacity (TLC) is difficult to measure, so it is useful to be able to predict TLC from other information. The following table shows the pre-transplant TLC of 32 recipients of heart-lung transplants, obtained by whole-body plethysmography,. and their age, sex and height (Otulana et al 1989).

ID	Age	Sex	Height	CTL (l)	ID	Age	Sex	Height	CTL (l)
1	35	F	149	3.40	17	30	F	172	6.30
2	11	F	138	3.41	18	21	F	163	6.55
3	12	M	148	3.80	19	21	F	164	6.60
4	16	F	156	3.90	20	20	M	189	6.62
5	32	F	152	4.00	21	34	M	182	6.89
6	16	F	157	4.10	22	43	M	184	6.90
7	14	F	165	4.46	23	35	M	174	7.00
8	16	M	152	4.55	24	39	M	177	7.20

9	35	F	177	4.83	25	43	M	183	7.30
10	33	F	158	5.10	26	37	M	175	7.65
11	40	F	166	5.44	27	32	M	173	7.80
12	28	F	165	5.50	28	24	M	173	7.90
13	23	F	160	5.73	29	20	F	162	8.05
14	52	M	178	5.77	30	25	M	180	8.10
15	46	F	169	5.80	31	22	M	173	8.70
16	29	M	173	6.00	32	25	M	171	9.45

(a) How well can an individual's lung capacity be predicted from a multiple regression model including age, sex and height?

(b) Compare the result just obtained with that derived from linear regression on height alone.

(c) Calculate 95% prediction interval from the linear regression on height for someone with average height.

(d) How could we investigate whether the relation between lung capacity and height is the same for males and females.

(e) Suppose that you are writing an article based on these analyses, what would you write in your "statistical method" section.