STATISTISK SENTRALBYRÅS HÅNDBØKER

Nr. 22 Oslo, 13. mars 1962

STATISTISK TESTING AV HYPOTESER VED REGRESJONSBEREGNINGER

HAROLD W. WATTS

TESTS OF COMPOSITE HYPOTHESES IN REGRESSION MODELS AND RELATED TOPICS

This paper is intended for limited circulation to stimulate discussion and critical comment. Since it is, in part, preliminary and tentative, references to this paper in publications should be cleared with the author.

Statistisk Sentralbyrå
Oslo, 1962

F o r o r d

Harold W. Watts, Assistant Professor of Economics ved Yale University, som for tiden oppholder seg i Norge med stipend fra Social Science Research Council, har utarbeidd dette kompendium i tilknytning til et uformelt seminar hvor en del av Byråets funksjonærer var med.

Kompendiet som behandler matematisk-statistiske emner, vil særlig være av interesse for de av Byråets funksjonærer som er fortrolige med DEUCE standardprogram for regresjonsberegninger. Forfatteren gjør utførlig rede for hvorledes en kan bruke beregningsresultatene til å teste hypoteser om regresjonskoeffisienter, ikke bare en av gangen, men også sammensatte hypoteser hvor et større antall koeffisienter testes under ett.

Byrået takker professor Watts for at han har vært så elskverdig å utarbeide dette kompendium.

Statistisk Sentralbyrå, Oslo, 13. mars 1962

Signy Arctander

Arne Amundsen

A u t h o r ' s n o t e

This paper elaborates and presents in a more orderly manner the topics discussed by the author at an informal seminar with several members of the staff of the Central Bureau of Statistics.

Oslo, March 13, 1962

Harold W. Watts

Introduction

The theoretical basis for tests of composite hypotheses is presented in most textbooks of mathematical statistics. As applied to the coefficients of a regression equation these results can be stated quite elegantly for the general class of linear hypotheses (i.e. where the n regression coefficients satisfy $k \leq m$ linear equations). The main problem is that the methods are not applied in many cases where they are clearly called for. Often a set of binary variables is introduced into a regression to represent some qualitative factor. If one asks whether this factor has a significant influence on the dependent variable the answer can only be given by a composite test. The errors estimated for the coefficients of individual binary variables are not very interesting in this situation. Again, one may be confident that either x and x^2 both appear in a regression or they are both absent. In this case also, a composite test is needed. Still another situation which require a composite test arises from hypotheses that some or all coefficients of an equation estimated from one sample are equal (in the vector sense) to a similar set of coefficients estimated from a different sample. Of course all these examples can be fitted into the framework of general linear hypotheses and pressed through the general test procedures. But it is often possible to define and calculate identical tests in terms of magnitudes that are both easier to understand heuristically and easier to calculate.

This note explains the simplifications which can be made for a large class of composite tests. While not as inclusive as the class of all linear hypotheses, it contains enough recurrent problems to be worth special attention.

The General Theory

First a brief review of the test of a general linear hypothesis. (This discussion depends heavily on A.M. Mood, Introduction to the Theory of Statistics, pp 301-307.)

The regression model is:

$$(1) \quad y = a_1x_1 + a_2x_2 + \ldots + a_nx_n + u \quad ,$$

$f(u) = N(0, \sigma_u)$. (Assume for the present that all variables are measured as deviations from their means.)

The null hypothesis is:

$$
\begin{aligned}
c_{11}a_1 + c_{12}a_2 + \cdots + c_{1n}a_n &= b_1^o \\
c_{21}a_1 + c_{22}a_2 + \cdots + c_{2n}a_n &= b_2^o \\
&\vdots \\
c_{k1}a_1 + c_{k2}a_2 + \cdots + c_{kn}a_n &= b_k^o \quad , \quad k \leq n \ .
\end{aligned} \tag{2}
$$

Or in matrix notation: $Ca = b^o$.

This null hypothesis can be viewed as a set of k linear constraints on the coefficients of (1) or they can be viewed as defining equations for a new set of k parameters (call them b_i, $i = 1, 2 \ldots k$) which are, by null hypothesis equal to b_i^o, $i = 1, 2 \ldots k$. If the constraints in (2) are ignored then the b's defined by $b = Ca$ are simply a particular transformation of the a's. If model (1) is estimated from a sample the estimates of the a's, $\hat{a}$ (a column vector), imply a set of estimates for the b's, call it $\hat{b} = C\hat{a}$. The test of the null hypothesis amounts to an evaluation of the probability of the deviations, $\hat{b} - b^o$, if the null hypothesis is true. Since the new parameters are linear functions of the a's, the estimated variance - covariance matrix for the $\hat{b}$'s can be derived simply from the one for the $\hat{a}$'s. Let X denote the $T \times n$ matrix of observed values of the x_i's and Y denote the $T \times 1$ vector of observed values of y (y and the x's are measured from their respective sample means). The variance - covariance matrix for the $\hat{a}$'s is estimated by $(X'X)^{-1}s_u^2$ and the estimated variance - covariance matrix for the $\hat{b}$'s is $\left[C(X'X)^{-1}C'\right]s_u^2$, where:

$$
s_u^2 = \Sigma \hat{u}^2/(T-n-1) = \frac{Y'Y - Y'X(X'X)^{-1}X'Y}{T - n - 1}
$$

The familiar, and classical, procedure for testing a hypothesis about a normally distributed statistic, say z, is to form the ratio: $(z - z^o)/s_z$ = "t", where z^o is the hypothesised value of z and s_z is the estimated standard error of z. This procedure yields a statistic which is distributed as "t" under the null hypothesis and can be tested against critical values from published tables. It is also the case that a "t" distributed variable when squared has the F distribution with one degree of freedom in the "numerator" and as many degrees of freedom in the denominator as has the

corresponding "t". The critical values of F in the one degree of freedom case can be seen in published tables to be equal to the square of the corresponding critical values of "t". If the statistic above is squared it can be written:

$$"t"^2 = F_1 = (z - z^o)\,(s_z^2)^{-1}\,(z - z^o).$$

In the composite test, of which the simple test above can be viewed as a limiting case, the F-test statistic can be written:

$$(3)\quad F_k = \frac{(\hat{b} - b^o)'\left[C(X'X)^{-1}C'S_u^2\right]^{-1}(\hat{b} - b^o)}{k}$$

By this argument the F-test is seen to be a straight-forward k-variable analog of the familiar "t"-test with loss of information about the signs of the deviations from hypothesis (as a consequence the critical values of F correspond to equivalent two-tailed "t" tests).

The same test statistic is more commonly written as:

$$(4)\quad F_k = \frac{\frac{1}{k}(\hat{b} - b^o)'\left[C(X'X)^{-1}C'\right]^{-1}(\hat{b} - b^o)}{S_u^2}$$

In this latter expression the quadratic form in the numerator measures the difference in the residual sum of squares (in the dimension of the dependent variable) between the unrestricted model (1) and the restricted model where the coefficients are required to satisfy (2). This portion of the sum of squares is divided by k, the number of degrees of freedom lost by satisfying (2). The resulting mean square is then compared in ratio with the mean square residual from the unrestricted equation. Looking at the statistic in this way brings out the analogy with the F-ratios used in more simple analysis of variance problems.

A Special Case

Consider the special case where the null hypothesis is simply that $a_i = 0$, $i = n-k+1, n-k+2 \ldots n$. In terms of the general linear hypothesis this amounts to specifying $b^o = 0$, and $C = [0, I_k]$, where I_k is the k-rowed identity matrix and is preceded in C by $n-k$ columns of k zeroes.

Straight-forward evaluation of F_k involves inverting a k x k matrix (the lower-right k x k partition of $(X'X)^{-1}$) and calculation of a quadratic form. Since that quadratic form measures, finally, the difference in residual sum of squares between the restricted and unrestricted models, why not evaluate that difference by re-estimating equation (1) with the last k variables left out? This crude-but-effective expedient yields the same sum of squares plus a bonus in the form of estimates of a_i, i = 1, 2 .... n-k, i.e. the estimated parameters where the null hypothesis is assumed to hold. The procedure for carrying out the test mentioned above can be summarized simply:

1) estimate the unrestricted model: $y = a_1x_1 + a_2x_2 + \ldots + a_nx_n + u$, and evaluate $\Sigma\hat{u}^2$ (sum of estimated squared residuals).

2) estimate the restricted model: $y = a'_1x_1 + a'_2x_2 + \ldots + a'_{n-k}x_{n-k} + u$, and evaluate $\Sigma\hat{\hat{u}}^2$.

3) evaluate $F_{k,T-n-1} = \dfrac{\Sigma\hat{\hat{u}}^2 - \Sigma\hat{u}^2}{\Sigma u^2} \cdot \dfrac{T-n-1}{k}$

If each of the sums in the expression for F is divided by Σy^2 (remember that y is measured from its mean) then F can be written simply as:

$$(5)\quad F = \frac{(1-R'^2)-(1-R^2)}{(1-R^2)} \cdot \frac{T-n-1}{k} = \frac{R^2-R'^2}{1-R^2} \cdot \frac{T-n-1}{k},$$

where R' and R are the multiple correlation coefficients in the restricted and the unrestricted model respectively. In some cases this form may be more convenient.

If k is very small relative to n the more general procedure may be more economical, particular if the work is to be done on a desk calculator. On the other hand if the estimation of the unrestricted model is carried out on a desk machine most calculating schemes permit evaluation of the sum of squared residuals at any stage of the one-variable-at-a-time reduction process. It is possible to have the same intermediate information provided by a high-speed computor since they usually use similar algorithms. But even without that feature the cost of re-estimation of the restricted equation is not very large, a major consideration is that the same program can be used for both the restricted and unrestricted equations. Moreover the lengthy calculation of moments does not have to be repeated.

The following mathematical argument establishes the equivalence between the general test procedure and the procedure suggested for the special case discussed above.

Let the matrix of observations on the x_i, X, be partitioned into $[X_1, X_2]$, where X_1 contains the first n - k variables and X_2 the last k. The vector of estimated coefficients, $\hat{a}$, is conformably partitioned into $\begin{bmatrix}\hat{a}_1\\ \hat{a}_2\end{bmatrix}$. Let Y denote the T x 1 vector of observed values of y (measured from the mean of y). The unrestricted a's are estimated by:

$$\begin{bmatrix}\hat{a}_1\\ \hat{a}_2\end{bmatrix} = \begin{bmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{bmatrix}^{-1} \cdot \begin{bmatrix}X_1'Y\\ X_2'Y\end{bmatrix}, \text{ and}$$

$$\Sigma\hat{u}^2 = Y'Y - [\hat{a}_1', \hat{a}_2'] \cdot \begin{bmatrix}X_1'Y\\ X_2'Y\end{bmatrix}$$

If $a_2 = 0$ is assumed to hold in the restricted equation then the last k variables can be ignored and the estimates for a_1 are simply:

$$\hat{\hat{a}}_1 = (X_1'X_1)^{-1} X_1'Y , \text{ and}$$

$$\Sigma\hat{\hat{u}}^2 = Y'Y - \hat{\hat{a}}_1'X_1'Y .$$

The difference, $\Sigma\hat{\hat{u}}'^2 - \Sigma\hat{u}^2$, is evidently measured by:

$$Q = (\hat{a}_1' - \hat{\hat{a}}_1') X_1'Y + a_2'X_2'Y .$$

Now let $M_1 = \left[I - X_1(X_1'X_1)^{-1} X_1'\right]$, (a T x T matrix) and

$$B = (X_1'X_1)^{-1} X_1'X_2 , \text{ (an n-k x k matrix).}$$

With these definitions it is easily verified that:

$$\begin{bmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{bmatrix}^{-1} = \begin{bmatrix}(X_1'X_1)^{-1} + B(X_2'M_1X_2)^{-1} B' & -B(X_2'M_1X_2)^{-1}\\ -(X_2'M_1X_2)^{-1} B' & (X_2'M_1X_2)^{-1}\end{bmatrix}.$$

Using this inverse the unrestricted estimates can be written:

$$\begin{bmatrix}\hat{a}_1\\ \hat{a}_2\end{bmatrix} = \begin{bmatrix}\hat{\hat{a}}_1 - B(X_2'M_1X_2)^{-1} (X_2'M_1Y)\\ (X_2'M_1X_2)^{-1} (X_2'M_1Y)\end{bmatrix} = \begin{bmatrix}\hat{\hat{a}}_1 - B\,\hat{a}_2\\ \hat{a}_2\end{bmatrix},$$

since $X_2'Y - B'X_1'Y = X_2'Y - X_2'X_1(X_1'X_1)^{-1} X_1'Y$

$= X_2'(I - X_1(X_1'X_1)^{-1} X_1') Y$

$= X_2' M_1 Y$

Consequently, $\hat{a}_1 - \hat{\hat{a}}_1 = - B\hat{a}_2$ and

$$Q = - \hat{a}_2'B'X_1'Y + \hat{a}_2'X_2'Y$$

$$= \hat{a}_2'(X_2'Y - B'X_1'Y)$$

$$= \hat{a}_2'(X_2'Y - X_2'X_1(X_1'X_1)^{-1} X_1'Y)$$

$$= \hat{a}_2'(X_2'M_1Y)$$

$$= \hat{a}_2'(X_2'M_1X_2)\, \hat{a}_2 \quad .$$

In this final form Q is shown to be equal to a quadratic form involving the estimated coefficients of the last k variables and the inverse of the lower right-hand $k \times k$ partition of the inverse matrix (cf. numerator of (4)).

Adaptation of the special case

The preceding section explained a simple device for evaluating the composite hypothesis that a set of coefficients are all equal to zero. That case is a common one although it is probably used more often than is justified. A slightly more general case can be handled almost as easily. Consider the null hypothesis:

$$a_i = b_i^o \quad , \quad i = n-k+1, n-k+2, \ldots, n.$$

I.e. the last k coefficients in the regression equation are hypothesised to be equal to a given set of numbers (b's). This case can be turned into the earlier one by transforming the dependent variable according to:

$$y_o = y - \sum_{i=n-k+1}^{n} (x_i b_i^o) \quad .$$

In terms of this transformed variable the equivalent hypothesis holds that the coefficients of the last k variables are equal to zero and the test procedures outlined above can be applied. The transformation can be made either before or after the moments are computed depending on convenience. (The dependent variable need not be transformed for the unrestricted equation since the sum of squared residuals will be the same for y as for y_o, but of course R^2 will be different.)

As another example consider the hypothesis that the last k coefficients are equal, i.e. :

$$a_{n-k+1} = a_{n-k+2} = \cdots\cdots = a_n .$$

For this hypothesis define a new variable:

$$x_{n+1} = \sum_{i=n-k+1}^{n} x_i .$$

Then compare the estimated residual sum of squares from the regression models:

$$y = \sum_{i=1}^{n} a_i x_i + u \qquad \text{(unrestricted)}$$

and

$$y = \sum_{i=1}^{n-k} a_i' x_i + a_{n+1}' x_{n+1} + u' \qquad \text{(restricted)}$$

It is trivially true that any general linear hypothesis can be turned into what is here called a special case by a suitable linear transformation of variables. The main point of the discussion above is that in many useful cases the transformation is quite simple and provides a practical alternative to the general procedure.

A Digression on "Constant Terms" in Regression Models

Regression analysis is commonly treated as a special branch of statistical analysis having to do with relations among variables. A logical desire to separate analysis of the relation between variables from the question of the absolute level of the variables has given rise to a convention of measuring variables from an origin chosen to coincide with the mean value of the variables involved. This convention is usually applied in the computing algorithms for obtaining estimates as well as in the theory. Typically

the "normal equations" are written in terms of the sums of squares and products of variables measured from their means. As a second step the so-called constant term is estimated by the additional condition that the regression surface must pass through the point representing the mean of all variables. This second step must be paralleled by a special procedure for obtaining the estimated sampling error of the constant term if that is needed.

This special treatment of one of the parameters in the model is not dictated by any mathematical or statistical necessity. If a "variable" which always takes the value 1 (this does stretch the use of language somewhat) is introduced into the regression equation having as its coefficient what is usually called the constant term, it is then obvious that the "constant" can be treated symmetrically with all other coefficients in the regression. The normal equations will now be written in terms of the sums of squares and products measured from zero as the origin and there will be one additional equation and also one additional term in each equation. If the new variable is denoted $x_0 \equiv 1$ (call it the identity variable) then all regression equations can be written in the form $\Sigma a_i x_i + u$. In a limiting case where $y = a_0 x_0 + u$, the estimate of a_0 is simply $\hat{a}_0 = \Sigma y / T = \bar{y}$. Moreover the estimated sampling error of $\hat{a}_0$ when ordinary regression formulae are used is:

$$S_{a_0} = \sqrt{\frac{\Sigma (y - a_0)^2}{T(T-1)}}$$, the ordinary expression for the sampling error of an estimated mean.

The impression is widespread that the constant term and its sampling distribution are intrinsically different from the other coefficients in a regression, and that no one is ever really interested in the constant term. In many cases one may be properly interested in the "slope" of a relationship to the exclusion of the "level" but there are cases where the reverse is true. Consider the model:

$$y = a + bx + u$$

where b is the parameter of primary interest and u is distributed with zero mean but a variance which is proportional to x^2 (u is hetero-scedastic). The model:

$$y/x = a(1/x) + b + u/x$$

is equivalent to the other model except that now the residual, u/x, is homo-scedastic and therefore the parameters can be estimated more efficiently. However b remains the parameter of interest and now it appears as the constant

term. Apparently even if one is only interested in slopes, he may sometimes need to know the sampling error of a nominal constant term.

Besides permitting symmetrical treatment of all parameters in the regression equation, the introduction of x_0 facilitates the development of regression analysis as a generalization from the case of a variable distributed with a constant mean to the case of a variable with a shifting mean. Alternatively the constant mean case can be treated as a special case of regression analysis.

What is more important, all the foregoing discussion of composite tests can be re-read dropping the assumption that all variables are measured from their means if the identity variable is added to the list of variables in the model. The constant term, a_0, is just like all the other parameters and can be tested (or not tested) jointly with other parameters. No special treatment is required. With the introduction of the identity variable it becomes easier to present, in the next section, some further adaptations of the testing procedure.

In situations where the sample means for the x's have a broader significance, as they would if the sample were drawn by some random process from a stable population, there is some sense in treating the "constant term" asymmetrically. Or, it may be more convenient to do so when, for example, the estimates are used to form conditional prediction intervals. It may be useful, in such cases, to transform the estimated equation of conditional expectation:

$$\hat{y} = \hat{a}_0 x_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \ldots + \hat{a}_n x_n \quad ,$$

into the alternate form:

$$\hat{y} = \bar{y} + \hat{a}_1(x_1 - \bar{x}_1) + \hat{a}_2(x_2 - \bar{x}_2) + \ldots + \hat{a}_n(x_n - \bar{x}_n) \quad ,$$

where $\bar{y}, \bar{x}_1, \bar{x}_2 \ldots \bar{x}_n$ are the sample means.

The variance - covariance among the $\hat{a}_i$, $i = 1, 2, \ldots, n$, are the same for both forms, the estimated variance of $\bar{y}$ is simply S_u^2/T, and the $\mathrm{Cov}(\bar{y}, \hat{a}_i) = 0$, $i = 1, 2, \ldots, n$. While this form may often be useful as a convenience, it should not be allowed to obscure the basic symmetry of all parameters in the equation of conditional expectation.

Some Further Adaptations of the Special Case of the Linear Hypothesis

Consider the problem of testing the hypothesis that the same regression model holds for two distinct sub-populations given a sample from each population. On the assumption that the error variance, σ_0^2, is the same in both populations the test can be carried out in a simple manner by adapting the techniques set out above. The "unrestricted" model for the combined sample (all coefficients are allowed to be different for the two sub-samples) can be written:

$$(6)\quad y = d_1 \sum_{i=0}^{n} a_{i1}x_i + d_2 \sum_{i=0}^{n} a_{i2}x_i + u,$$

where d_1 and d_2 are binary variables defined according to:

d_1	d_2	take values
1	0	for members of the first sub-sample, and
0	1	for members of the second sub-sample.

(Note that the indentity variable, x_0 is now included in the model and so the variables no longer need be measured as deviations from means.)

A trivial transformation of the model above provides:

$$y = \sum_{i=0}^{n} a_{i1}x_i + d_2 \sum_{i=0}^{n} (a_{i2} - a_{i1})x_i + u .$$

In this form the null hypothesis can be stated as:

$$(a_{i2} - a_{i1}) = b_i = 0, \quad i = 0, 1, 2, \ldots, n .$$

It is now obvious that the "restricted" model can be estimated from the combined sample by fitting:

$$(7)\quad y = \sum_{i=0}^{n} a_i'x_i + u' .$$

The "restricted" sum of squares is $\Sigma\hat{u}'^2$. The "unrestricted" sum of squares, $\Sigma\hat{u}^2$, can be evaluated most easily by fitting:

$$y = \sum_{i=0}^{n} a_{i1}x_i + u_1 \quad \text{to the first sub-sample, and}$$

$$y = \sum_{i=0}^{n} a_{i2}x_i + u_2 \quad \text{to the second sub-sample.}$$

Then $\Sigma\hat{u}^2 = \Sigma\hat{u}_1^2 + \Sigma\hat{u}_2^2$.

Recapitulating, the procedure for testing the difference between two sample regressions is:

1. Fit the model $y = \sum_{i=0}^{n} a_i x_i + u$ to:

   a. the T_1 observations from sub-population 1 (this yields $\Sigma\hat{u}_1^2$ with T_1-n-1 degrees of freedom)

   b. the T_2 observations from sub-population 2 (this yields $\Sigma\hat{u}_2^2$ with T_2-n-1 degrees of freedom)

   c. the T_1+T_2 observations in the total sample (this yields $\Sigma\hat{u}'^2$ with T_1+T_2-n-1 degrees of freedom).

2. Form the test statistic:

$$F = \frac{\Sigma\hat{u}'^2 - \Sigma\hat{u}_1^2 - \Sigma\hat{u}_2^2}{\Sigma\hat{u}_1^2 + \Sigma\hat{u}_2^2} \cdot \frac{T_1 + T_2 - 2n - 2}{n + 1}$$

(Note that for the limiting case of $n = 0$ this expression is equal to the square of the "t" test statistic for the difference between two sample means given that they come from populations with equal variance.)

It is clear how this procedure can be extended to cases where more than two population regressions are compared.

The restricted and unrestricted models above represent two extremes, either all $n+1$ coefficients are the same or all are different. There are a large number of intermediate models between these extremes in which some coefficients are allowed to differ and some are not. It will be instructive to examine one of these intermediate models. A commonly encountered one is where all the "slopes" $(a_i, i = 1, 2, \ldots, n)$ are required to be the same for both sub-populations but the intercept (a_0) is allowed to differ, i.e.:

$$(8) \qquad y = d_1 a_{01} x_0 + d_2 a_{02} x_2 + \sum_{i=1}^{n} a_i x_i + u'' .$$

Rewritten in the form:

$$(8a) \qquad y = a_{01} x_0 + (a_{02} - a_{01}) d_2 x_0 + \sum_{i=1}^{n} a_i x_i + u''$$

the model is easily recognized as the case of a "dummy" or binary variable providing a shift in the level of the regression surface for members of population 2. Given $\Sigma\hat{u}''^2$ evaluated from this model the following hypotheses can be tested:

1. The hypothesis that the "intercepts" are different under the assumption that the "slopes" are all the same can be tested by a comparison between $\Sigma\hat{u}''^2$ and $\Sigma\hat{u}'^2$ (i.e. the hypothesis that $(a_{02} - a_{01}) = 0$ which can also be tested by a simple "t" test on the coefficient of the binary variable in (8a)).

2. The hypothesis that all "slopes" are different under the assumption that the "intercepts" are different can be tested by a comparison between $\Sigma\hat{u}''^2$ and $\Sigma\hat{u}^2$.

Each of the intermediate models can be compared in a similar way with the extreme models or with other intermediate models. Each such comparison involved a different hypothesis and implies a particular choice of restricted and unrestricted sub-spaces in the total parameter space. It is not the case that any pair of models formed by placing restrictions on (6) can be properly compared. To be valid the more restricted model must contain all the restrictions present in the less restricted model plus some additional ones (i.e. the more restricted sub-space must be wholly contained in the less restricted sub-space). There are computational shortcuts available for estimation of the intermediate models. These shortcuts are particularly helpful when more than 2 sub-populations are being compared. For a thorough discussion of these shortcuts see "On a Class of Regressions using Binary or 'Dummy' variables" by H. Watts (Memorandum of the Oslo University Institute of Economics).

On the Use of R^2 in Regression Analysis

Most elementary statistical textbooks, in a chapter on regression and correlation, define a statistic R, (or R^2), which is called the coefficient of multiple correlation (determination). One formula for this statistic is:

$$R^2_{y \cdot x_1, x_2, \ldots, x_n} = 1 - \frac{\Sigma (y - \hat{a}_0 x_0 - \hat{a}_1 x_1 - \ldots - \hat{a}_n x_n)^2}{\Sigma (y - \bar{y})^2}$$

The corresponding R is simply the positive square root of this expression. A common locution for explaining the meaning of R^2 is: "the proportion of the variance of y which is 'explained' by the regression on the x's".

On the other hand, a model of normal linear regression is completely specified by the coefficients (a_i's) in the equation of conditional expectation,

$E(y) = \Sigma a_i x_i$, together with the conditional variance, σ_u^2,. It is evident that R or R_2^2 is not necessary for specifying a regression model. Indeed it is impossible to associate a value of R^2 with a specific regression model for the simple reason that, in the absence of information about the distribution of the x's, the unconditional or marginal variance of y cannot be inferred. R^2 can be made as large (small) as desired by choosing x's which make the sample variance of y sufficiently large (small).

Even with a given set of values for the x's (or with a given distribution), trivial changes in the statement of a regression model can effect the value of R^2. Consider the models:

$$\text{a.} \quad y = a_0 x_0 + a_1 x_1 + u \,, \quad f(u) : N(0,\sigma) \,, \quad \text{and}$$

$$\text{b.} \quad y + kx_1 = a_0 x_0 + (a_1 + k)x_1 + u \,, \quad f(u) : N(0,\sigma)$$

For all intents and purposes the two models are equivalent for any specified value of k. With a given sample the two models yield identical estimates of the parameters a_0, a_1, and σ. Either can be used for forming conditional predictions of y. However by choosing $|k|$ sufficiently large, R^2 can be brought arbitrarily close to 1. Similarly, as k is brought close to a_1, R^2 approaches 0.

It can be protested that the R^2's are not comparable between the two models just discussed and indeed they are not since the "dependent" variable is not the same in the two cases. But the non-comparability has not been recognized widely enough. Students, and an occasional scholar, may prefer a regression which sets consumption as a linear function of income over an equivalent one which places saving (≡ income - consumption) as a linear function of income simply because the former typically provides a higher R^2.

The foregoing remarks would almost justify a crash program to stamp out the use of R^2 in regression analysis. Short of taking that drastic step, it is possible to sharpen this widely used tool. R^2 can be interpreted as an indicator, or index of the closeness of fit of a regression. As an index it has the desirable property of taking values between zero and one. The extremes correspond to explanation of none or all of the variation in the dependent variable. Like most indices, it does not permit valid comparisons unless the base of the index is comparable. The base in the case of R^2 is the sample variation of the dependent variable; different R^2's are comparable if the base is held constant. Within a given sample of observations on y and $x_1, x_2, x_3, \ldots, x_n$, many regression models can be estimated. Of these, the set

of models which use y as the dependent variable have, by definition, the same unconditional variance for the dependent variable and the R^2's for these models are therefore comparable. The same statement can be made for the set of models which use $\log(y)$ (or $1/y$, or $y-x_2/2$). If one wishes to compare two models which use different dependent variables, say y and $\log(y)$, then he cannot rely on the R^2's as usually computed. In this particular case he must first decide whether he wants to compare the two regressions in terms of their absolute or percentage deviations. Whichever comparison is chosen one set of residuals must be transformed into the other basis and then the mean-square error (or some other criterion) can be evaluated for both and compared.

A less strict notion of comparability might be invoked to permit comparisons between samples. It is still essential that the same dependent variable enter the base but the sample variance of that variable need not be exactly the same. The sample variance could be different because of sampling variation in the u's even if the values for the set of x's is the same for both samples. A more common case is where the two samples are obtained from the same population by the same sampling procedure and where there is a multivariate distribution of the x's (with finite variance) in that population. In this case the difference in the sample variance of y on account of different x's can again be attributed to sampling and the R^2's probably remain "comparable enough" for rough comparisons. It is in this sense that continued exposure to a certain kind of sample and basic model or set of models may establish norms for evaluation of additional R^2's of the same type. The preceding discussion should be sufficient to emphasize the relative nature of such norms.

The notion of comparability between R^2's should be refined further in the case where a single sample is involved. The sample consists of observations on y and $x_0, x_1, \ldots, x_n$. Suppose that the following three alternative regression models are estimated from the sample.

$$(9a) \qquad y = \sum_{i=0}^{n} a_i x_i + u_1$$

$$(9b) \qquad y = \sum_{i=0}^{k} a_i x_i + u_2 , \quad k < n , \text{ and}$$

$$(9c) \qquad y = a_0 x_0 + \sum_{i=k+1}^{n} a_i x_i + u_3 .$$

The R^2's associated with the three models in the given sample can be denoted R_1^2, R_2^2, and R_3^2 respectively.

R_1^2 is necessarily at least as large as the larger of R_2^2 and R_3^2. If R^2 is interpreted as an index of fit then model (9a) will never produce a poorer fit than (9b) or (9c) and will usually fit better. It is possible to test whether the difference between R_1^2 and, say, R_2^2 is larger than can be attributed to chance. That is equivalent to testing the hypothesis that the last $n-k$ of the x's make no net contribution to the explanation of y or that their coefficients are all equal to zero. The comparison of R_1^2 and R_3^2 implicitly refers to a similar test on the variables left out of (9b). What about the comparison between R_2^2 and R_3^2? It meets the requirements for comparability that were set out earlier but it does not correspond to any linear hypothesis about the regression coefficients. Yet there is some meaning to the statement that (9b) fits better than (9c) if $R_2^2 > R_3^2$. Probably one would prefer to use (9b) rather than (9c) for making predictions about y if he were faced with a choice between the two and could not use (9a). But there is not, at least at present, a theoretical basis for making a test based on that comparison.

Multiple-Partial Correlation Coefficients

In line with the discussion of the proper use of R^2 in regression analysis, a recent statistical innovation should be introduced.[1] Partial correlation coefficients have long been available for measuring the degree of association between two variables considered net of the relation each may have with one or more other variables. Thus $r^2_{y\cdot x_1/x_2x_3,\ldots,x_n}$ can be described as the proportion of the residual variation of y after regression on $x_2, x_3, \ldots, x_n$ which is "explained" by the residual of x_1 after regression on $x_2, x_3, \ldots, x_n$. The ordinary partial correlation can be defined by:

$$R^2_{y\cdot x_1x_2\cdots x_n} = R^2_{y\cdot x_2x_3\cdots x_n} + r^2_{y\cdot x_1/x_2x_3\cdots x_n}\left(1 - R^2_{y\cdot x_2x_3\cdots x_n}\right)$$

or

$$r^2_{y\cdot x_1/x_2x_3\cdots x_n} = \frac{R^2_{y\cdot x_1x_2\cdots x_n} - R^2_{y\cdot x_2x_3\cdots x_n}}{1 - R^2_{y\cdot x_2x_3\cdots x_n}}$$

---

1) See "Partial Trace Correlation Theory", Cowles Foundation Discussion Paper No. 97 (Mimeographed).

By simple extension define:

$$(10) \qquad R^2_{y\cdot x_1x_2\cdots x_k/x_{k+1}x_{k+2}\cdots x_n} = \frac{R^2_{y\cdot x_1x_2\cdots x_n} - R^2_{y\cdot x_{k+1}x_{k+2}\cdots x_n}}{1 - R^2_{y\cdot x_{k+1}x_{k+2}\cdots x_n}}$$

This statistic can be called the Multiple-Partial Correlation (squared). It bears the same relation to the ordinary partial correlation coefficient as the multiple correlation coefficient bears to the simple correlation coefficient. In terms of the R^2's for the three models discussed in the preceding section two multiple-partial correlations can be calculated. They are:

$$R^2_{y\cdot x_1x_2\cdots x_k/x_{k+1}x_{k+2}\cdots x_n} = (R^2_1 - R^2_3)/(1 - R^2_3) = MPR^2_{b|c}$$

$$R^2_{y\cdot x_{k+1}x_{k+2}\cdots x_n/x_1x_2\cdots x_k} = (R^2_1 - R^2_2)/(1 - R^2_2) = MPR^2_{c|b}$$

The discussion of R^2 has implicitly assumed that the identity variable, x_0, "belongs" in the regression equation. That assumption is typically made when R^2's are defined for regression models. If, in line with the earlier discussion of "constant terms", the inclusion of x_0 is to be regarded as an open question then the statistic that is usually called R^2 should be written as a multiple-partial correlation, i.e. as:

$$R^2_{y\cdot x_1x_2\cdots x_n/x_0} = \text{(by convention)} \quad R^2_{y\cdot x_1x_2\cdots x_n}.$$

Also x_0 should be added to the list of variables "partialled out" in the multiple-partial correlations defined above.

It is probably unwise to propose an alteration in the notation for R^2 at this time, but the fact that the conventional R^2 assumes that x_0 has been partialled out should be kept in mind. This is of prime importance whenever models that do not include x_0 are used. In those cases the sum of squares of y around zero may be used as the base of R^2 instead of the sum of squares around the mean.[1)]

A test of the hypothesis that a multiple-partial correlation coefficient is equal to zero corresponds to a composite hypothesis that a subset of

---

1) Estimated residuals in models without constant terms will not, in general, have a zero mean. Nevertheless, Σu^2 (not $\Sigma u^2 - T\bar{u}^2$) is the proper sum to compare with Σy^2.

regression coefficients is equal to zero. Thus the hypothesis that $MPR^2_{b|c} = 0$ corresponds to the hypothesis that $a_i = 0, 1, 2, \ldots, k$. It has been shown earlier that this hypothesis can be tested by computing the test statistic:

$$F_{k,T-n-1} = \frac{R_1^2 - R_2^2}{1 - R_1^2} \cdot \frac{T - n - 1}{k} \quad , \quad (\text{cf. } (5)).$$

By virtue of the definition of the multiple-partial correlation that statistic can as well be written:

$$F_{k,T-n-1} = \frac{MPR^2_{b|c}}{1 - MPR^2_{b|c}} \cdot \frac{T - n - 1}{k} \quad .$$

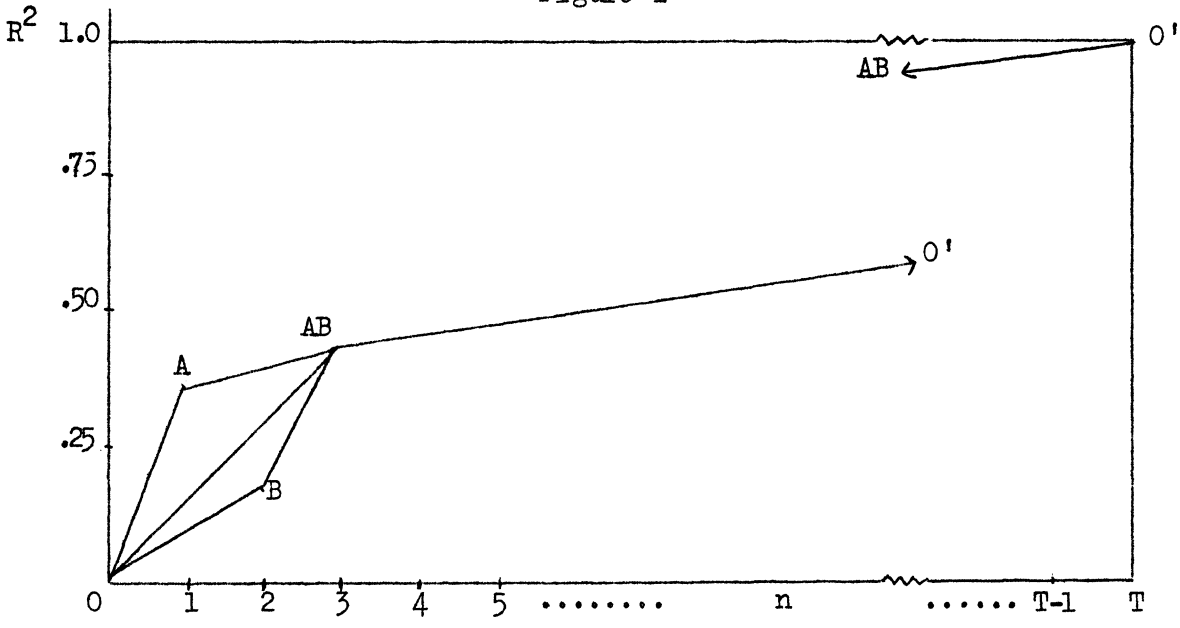
It should be re-emphasized that the remarks about R^2 have relevance only in terms of the basic regression model where the only random variable is the "residual", u. Since that model has nothing to say about the distribution of the x's that appear in the equation for the expected value of y, nothing can be inferred about the unconditional variance of y in the population. Consequently R^2 (multiple or multiple-partial) has no meaning within the framework of the model. Nevertheless it is possible to define R^2 as a simple statistic which describes a feature of a <u>sample</u> which does have a specific set of x_i values associated with it. It is questionable whether they <u>should</u> be computed in applications of the regression model. But if they are, and it is common practice to do so, then it is essential that they be used sensibly. R^2 can be interpreted as a descriptive index of fit or as the complement of a standardized measure of the residual variance. As such, if the base of the index or the basis of standardization is kept in view, they can be of some use.

The practice of computing R^2's probably derives from a superficially similar but distinct model in which y and the x's are jointly distributed according to some multi-variate density function. In such models the R^2's have more theoretical significance. Indeed, if the joint distribution is normal there is a direct theoretical counterpart of sample R^2.

Test - O - Grams

Within the constraints placed on the use of R^2 in regression analysis, a graphical method is available for representing a group of several comparable R^2's. In a rectangular diagram let the vertical axis be labelled R^2 and graduated from 0 to 1. The horizontal axis is labeled "degrees of freedom" (n) and is graduated from 0 to T (see figure 1). On this diagram plot points with coordinates (n,R^2) where R^2 is the sample R^2 for a particular model and n is the number of coefficients estimated in the model. For

Figure 1



example, point A in the diagram might represent the model:

$$y = a_0x_0 + u \qquad (x_0 \text{ is the identity variable so } R_A^2 = 1 - \Sigma(y-\bar{y})^2/\Sigma y^2 = T\bar{y}^2/\Sigma y^2),$$

B might represent:

$$y = a_1x_1 + a_2x_2 + u' \quad (R_B^2 = 1 - (\Sigma u'^2/\Sigma y^2)), \text{ and}$$

AB might represent:

$$y = a_0''x_0 + a_1''x_1 + a_2''x_2 + u'' \quad (R_{AB}^2 = 1 - (\Sigma u''^2/\Sigma y^2)).$$

Connecting lines may be drawn between points if the model represented by the point farther to the left can be viewed as a linearly

restricted version of the model represented by the point to the right. The slope of the line is then a measure of the "cost" of the restrictions in terms of R^2 per degree of freedom. Thus the diagram shows lines connecting the point AB with A, B, and the origin (the origin corresponds to the model $y = u$). There are also lines to join A and B to the origin. There is, however, no line between A and B.

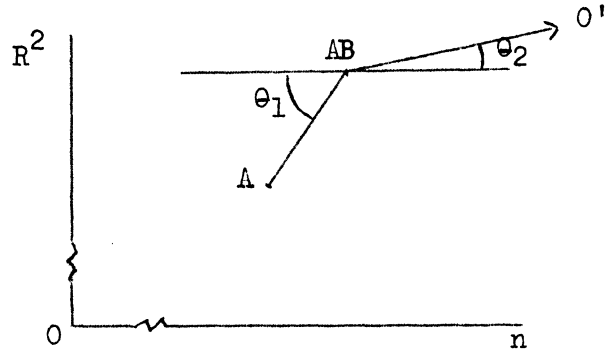
Multiple-partial correlations can be inferred roughly from the diagram by comparing the vertical distance between two points (if they can be validly joined by a connecting line or lines) with the vertical distance between the lower of the points and the upper boundary of the diagram (the line $R^2 = 1$).

F-Ratio's can also be evaluated on the diagram by a comparison of slopes. The F-Ratio for testing the hypothesis that all three coefficients in the third model (AB) are equal to zero is:

$$F_{3,T-3} = \frac{R^2_{AB}/3}{(1-R^2_{AB})/(T-3)}$$

That quantity can be represented as the ratio of the slope of the line from the origin, 0, to AB to the slope of the line from AB to 0' in the upper right-hand corner. Similarly the test of the hypothesis that $a''_1 = a''_2 = 0$ regardless of the value of a''_0 can be evaluated by the ratio of the slope of the line A, AB to the slope of the line AB, 0'. In terms of the angles shown in figure 2, $F = \tan(\theta_1)/\tan(\theta_2)$

Figure 2



In the example, Σy^2 was the base for all the R^2's, but the base could have been $\Sigma (y - \bar{y})^2$ and the ordinates of the points A, B, and AB could have been, respectively:

$$R^2_{y\cdot x_1/x_0} \quad , \quad R^2_{y\cdot x_2x_3/x_0} \quad , \text{ and } \quad R^2_{y\cdot x_1x_2x_3/x_0} \ .$$

This is the more conventional case where the fact that the R^2's are partial with respect to x_0 is ignored. (In this case the horizontal axis would only extend to $T - 1$.)

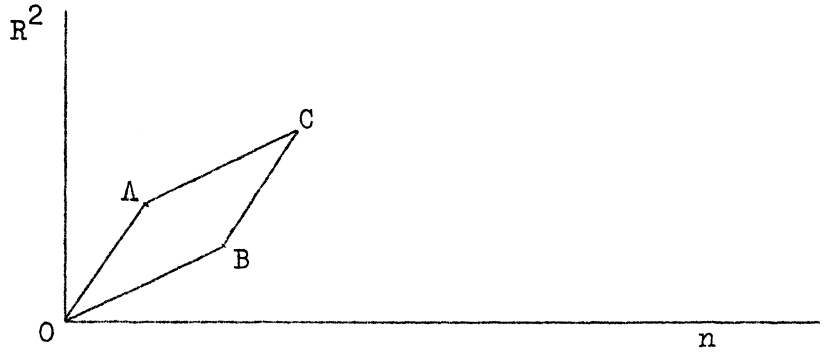
As another alternative the ordinates of A, B, and C could have represented, respectively:

$$R^2_{y\cdot x_1/x_0x_4x_5} \quad , \quad R^2_{y\cdot x_2x_3/x_0x_4x_5} \quad , \text{ and } \quad R^2_{y\cdot x_1x_2x_3/x_0x_4x_5} \ .$$

This would imply using the residuals from a regression on x_0, x_4, and x_5 as a base. (The degrees of freedom would only extend to $T - 3$ in this case.)

The pattern of points in a test-o-gram can provide some clues about the degree and nature of interdependence among the estimated coefficients. Let point C in the following diagrams represent some multiple regression model (call it model C). Points A and B represent models which contain non-overlapping subsets of the variables in C and which, taken together, exhaust the list of variables in C (call the subsets x_a and x_b. It is clear that point C must lie above, or on a par with, the higher of the points A and B. If the estimates of the coefficients of x_a and x_b are independent in model C, then the test-o-gram representation of the three models will show a perfect parallelogram as in figure 3.

Figure 3



Where the subsets of coefficients are not independent the pattern may resemble figure 4 or figure 5. In figure 4, C is much higher than the combined heights of A and B. In figure 5, C is not much higher than either A or B. These examples correspond to the multi-variable extensions of positive

and negative covariances between pairs of estimated coefficients. In the multi-variable context it is probably more precise to say that x_a and x_b

Figure 4

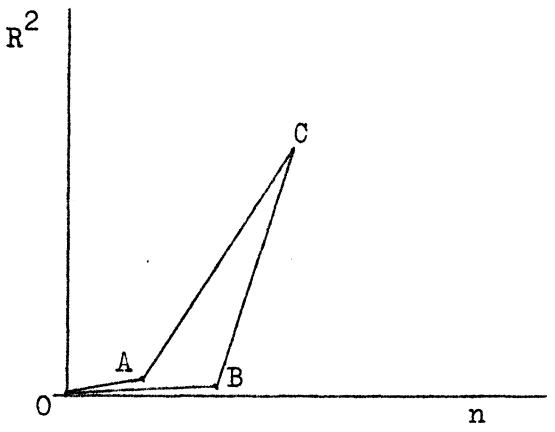
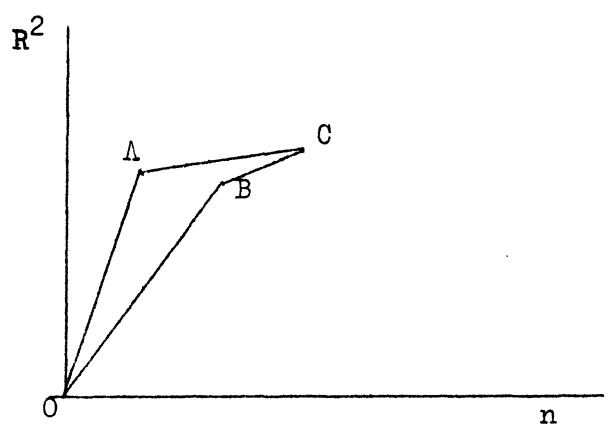


Figure 5



complement each other if the pattern resembles figure 4, and that they compete with each other in "explaining" y if the pattern is more like figure 5.

General purpose test-o-gram forms can be constructed by changing the horizontal axis to measure degrees-of-freedom-used as a proportion of the total degrees of freedom (it could be labeled n/T). Several standard forms with different ranges for n/T and R^2 would avoid the problem of finding all points in a very small segment of the diagram. For convenience a few of the rays which pass through 0' (i.e. the lines generated by $R^2 = 1 - a + an/T$ as a is varied) can be ruled on the diagram to facilitate graphical evaluation of F-ratios.

Use of the test-o-gram is entirely a matter of convenience, it adds nothing to what can be learned from a simple list of residual sums of squares together with corresponding degrees of freedom. But it can convey that information efficiently and in a form that permits simultaneous consideration of several comparable models.

Liste over publikasjoner i serien Statistisk Sentralbyrås Håndbøker (SSH)

Nr. 1 Regler for publikasjonenes utstyr m.v. i serien Statistisk Sentralbyrås Håndbøker

" 2 Veiledning for nye assistenter

" 3 Regler for maskinskriving i Statistisk Sentralbyrå

" 4 Innføring i maskinregning. Hefte 1. Addisjonsmaskiner

" 5 Innføring i maskinregning. Hefte 2. Kalkulasjonsmaskiner

" 6 Regler for utstyr m.v. for publikasjoner i serien Norges offisielle statistikk (NOS) og Samfunnsøkonomiske studier (SØS) og publikasjonen Statistiske meldinger

" 7 Retningslinjer for det skjematekniske arbeid i Byrået

" 8 Framlegg til nordisk statistisk terminologi

" 9 Standard for næringsgruppering i offentlig norsk statistikk

" 10 Hjemmel for innkreving av oppgaver

" 11 Kurs i hullkortmaskiner

" 12 Adresseliste over de kommunale folkeregistre

" 13 Standard for handelsområder

" 14 Innføring i DEUCE

" 15 Programmering for DEUCE. Første hefte

" 16 Alfasystemet. Et lettkodingssystem for DEUCE

" 17 Håndbok for DEUCE-operatører

" 18 Programmering for DEUCE. Annet hefte

" 19 Varenomenklatur for industristatistikken

" 20 Regler for publiseringsarbeidet m.v. i Statistisk Sentralbyrå

" 21 Håndbok for 1401-programmerere og -operatører

" 22 Statistisk testing av hypoteser ved regresjonsberegninger