A TALE OF TWO REGRESSIONS

by

Allan R. Sampson¹

FSU Statistics Report M263
ONR Technical Report No. 75

July, 1973

The Florida State University
Department of Statistics
Tallahassee, Florida

¹This research was in part supported by ONR Grant No. N00014-67-A-0235-0006. Reproduction in whole or in part is permitted for any purpose of the United States Government.
Abstract

A Tale of Two Regressions

by

Allan R. Sampson
Department of Statistics
Florida State University

This expository paper details the effects of the choice of underlying models in the statistical analysis of regression. The two models considered are the cases when all the data is jointly normally distributed and when the independent variables are assumed to be constants. The standard results for both these models are related to each other conditionally in a unified manner and an example is given where new results in one model can be obtained from old results in the other model. Particular emphasis is devoted to demonstrating differences that arise in power function considerations.
A Tale of Two Regressions

by

Allan R. Sampson
Department of Statistics
Florida State University

1. Summary and Background.

When analyzing data by what is conventionally termed "regression analysis", we must very often make an implicit choice. We must decide at least for ourselves whether to view the independent variables as constants or as realizations of random variables. However, we are aware that our decision matters little because the analysis is "essentially" equivalent. This paper deals with the conditional relationships between these analyses.

The concept of this equivalence is quite old. Bartlett (1933) in a remarkable paper established some of the basic material in this area. Unfortunately, his presentation was somewhat hampered as he did all of the necessary matrix analysis elementwise. Rao, following an approach similar to Bartlett, gave reference to this equivalence in both his text (1952) and a later paper (1959). More recently Dempster (1969) considered this parallelism of analyses from a different point of view. Certainly, there are numerous other authors who make at least passing reference to this problem.

There is also a parallel growth of literature in another area which can be best termed as philosophy of regression. Some early and basic works are by Kendall (1951), (1952), Berkson (1950),

---

1This research was in part supported by ONR Grant No. N00014-67-A-0235-0006. Reproduction in whole or in part is permitted for any purpose of the United State Government.
Windsor (1946), and Eisenhart (1947). The scope of our paper is much more restricted than that of these papers, being neither as far reaching nor as philosophical.

Our goal is to compile the previously established results about this conditional equivalence and give a unified and rigorous presentation of this material. In doing so, we pay special attention to the effect of the choice of analysis and to the capability of deriving results in one scheme from the other, particularly in deriving more complicated results from simpler ones.

We discuss the underlying models that give rise to the two viewpoints of analysis, restricting our attention to the cases where the randomness of the model has a normal distribution. We re-establish the theorems that tie together the scheme of "constant independent variables", which we simply call regression analysis, and the scheme of random independent variables (or realizations thereof), which for reasons to be seen we call multivariate analysis of regression. The dissimilitudes are also discussed with particular attention paid to the power function in the hypothesis testing situations. Finally we show how to use some of the comparative theorems to derive new results in multivariate analysis of regression from known simpler results in regression analysis.

2. The Underlying Models.

Often for experiments that result in a set of data vectors, a
linear relationship is sought among the variables. This linear relationship can be used for inference and prediction.

For example, we may wish to study the relationship between a student's college grade point average (GPA) and his high school GPA and "college board" scores. Typically the high school GPA and "college board" scores are considered to be independent variables; then the college GPA is called the dependent variable.

The regression analysis treatment of such data requires viewing the independent variables as fixed and the dependent variable as being random (or as being the realization of a random variable once the data is collected). On the other hand, the multivariate analysis of regression treatment views the triplet (college GPA, high school GPA, "college board" scores) as a trivariate random variable (or again as a realization, thereof, once the data is collected).

In this example, to treat the independent variables as fixed and the dependent variables as random appears debatable. There seems to be no reason to impose a qualitative difference between college GPA and high school GPA -- they are both the same kind of variable. While this lack of qualitative difference apparently dictates use of multivariate analysis of regression, data of this sort is often approached using regression analysis (i.e., see Draper and Smith (1966)). In fact, it is common to see either type of analysis used to analyze such data.

To this end, a restriction is imposed on the independent variables.
We want the independent variables not to be predetermined; that is, we cannot beforehand choose the values at which to observe the dependent variables. If this were able to be done, it would make no sense to view the independent variables as the realizations of a (non-trivial) random variable.

We now proceed, after establishing notation, to give the formal statistical models associated with regression analysis and multivariate analysis of regression. Denote the continuous random variable corresponding to the independent variable by $X$, $X$ being $p$-dimensional. (Kendall terms this the predicated variate.) Then $x$ is a realization of $X$ (called the predicated variable by Kendall). Similarly the random variable corresponding to the dependent variable and its realization are denoted $Y$ and $y$, respectively. (Kendall designates these the unpredicated variate and variable, respectively.) Assume we have a sample of size $n$, $Y_1$, $X_1$, ..., $Y_n$, $X_n$ and let $Z_1 = (Y_1, X_1)$. The associated realizations of these random variables are $y_1$, $x_1$, ..., $y_n$, $x_n$ and $z_1$, ..., $z_n$. Further define $\hat{X} = (x_1, ..., x_n)$, $\hat{Y} = (y_1, ..., y_n)$, and their realizations $\hat{x} = (x_1, ..., x_n)$, $\hat{y} = (y_1, ..., y_n)$. (Thus, the matrix $X$ has dimension $n \times p$.)

As usual $N(\mu, \Sigma)$ denotes the law of the multivariate normal random variable with mean $\mu$ and non-singular covariance matrix $\Sigma; n(\underline{x}; \underline{\mu}, \Sigma)$ denotes the density corresponding to $N(\mu, \Sigma)$. In general, $\mathcal{N}(U)$ denotes
the law of a random variable $U$ and $\mathcal{L}(U|V = v)$ denotes the conditional law of $U$ given $V = v$. (For the sake of clarity, the conditioning random variable's value is almost always included.)

**Model A (Regression Analysis):**

$$Y = X\beta + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2 I)$. Both $\sigma^2$ and $\beta$ are the standard parameters of interest.

**Model B (Multivariate Analysis of Regression):**

for $1 \leq i \leq n$,

$$Z_i \sim \text{i.i.d. } N(0, \Sigma).$$

For Model B, the standard parameters equivalent to $\beta$ and $\sigma^2$ of Model A are

$$\sum^{-1}_{22} \sum_{21} \text{ and } \sum_{11} - \sum_{12} \sum^{-1}_{22} \sum_{21},$$

where

$$\sum = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}$$

(1)

and $\sum_{21}$ is $p \times 1$.

The vector $\sum^{-1}_{22} \sum_{21}$ is the appropriate parameter for many reasons. The common justification is that for $1 \leq i \leq n$,

$$E(Y_i - X_i \gamma)^2$$

is minimized for $\gamma = \sum^{-1}_{22} \sum_{21}$, so that $X_i \sum^{-1}_{22} \sum_{21}$ is the best linear predictor of $Y_i$ (of course, it is the best predictor).
in the sense of minimizing squared error loss. When we speak of "regression coefficients", we mean either $\beta$ or $\sum_{22}^{-1} \sum_{21}$ depending on the underlying model.

Furthermore, under Model B

\[(2) \quad \tilde{N}(Y|\tilde{X} = X) = N(X \sum_{22}^{-1} \sum_{21}, (\sum_{11} - \sum_{12} \sum_{22}^{-1} \sum_{21})I),\]

where $I$ is the $n$-dimensional identity matrix. This distribution theory result is often interpreted by saying that Model A is a "conditional version" of Model B, that is, conditional on knowing the independent variables. Moreover, from (2) it is clear why we want to identify $\sum_{11} - \sum_{12} \sum_{22}^{-1} \sum_{21}$ with $\sigma^2$ of Model A.

For explanatory and notational ease, we have made the assumption in these models that the "regression plane" passes through the origin. In Model A, this assumption is equivalent to assuming $\tilde{X}$ cannot be forced to take predetermined values and, in particular, no column of $X$ can be all ones (almost surely). In Model B, this assumption is essentially equivalent to assuming $E \tilde{Z}_1 = 0$. (It can be shown that in the conditional probability statements which follow that this assumption is not restrictive because $\tilde{Z}$ and the maximum likelihood estimator of $\sum$ are independent under Model B and a similar result holds for Model A.) Note that from the lack of predetermination in $\tilde{X}$, or, equivalently, from the fact that the columns of $X$ are non-singular multivariate normal
vectors, we can assume that \( X'X \) is non-singular (almost surely).

3. Estimation.

In this section it is shown that the maximum likelihood (ML) estimates obtained under both models are the same while the ML estimators necessarily differ, being defined on different sample spaces. Here we make the usual distinction between estimate and estimator. Simply an estimator is a function of random variables; and an estimate is that number obtained by evaluating the corresponding estimator at the realizations of the random variables.

Let \( \mathcal{M}_{n,p} \) denote the set of \( n \times p \) matrices of full rank and \( \mathcal{V}_m \) denote the set of \( m \) dimensional vectors. Define for \( n \geq p \),

\[
B: \mathcal{M}_{n,p} \times \mathcal{V}_n \rightarrow \mathcal{V}_p \quad \text{and} \quad V: \mathcal{M}_{n,p} \times \mathcal{V}_n \rightarrow \mathcal{V}_1
\]

by

\[
B(A,b) = (A'A)^{-1} A'b,
\]

\[
V(A,b) = b'b - B'(A,b)(A'A)B(A,b),
\]

where \( A \in \mathcal{M}_{n,p} \) and \( b \in \mathcal{V}_n \).

Under Model A, the ML estimator \( \hat{\beta}_A \) of \( \beta \) is given by

\[
\hat{\beta}_A = (X'X)^{-1} X'Y = E(X, Y)
\]

and the ML estimator \( \hat{\sigma}^2_A \) of \( \sigma^2 \) by
(6) \[ \hat{\sigma}^2_A = n^{-1}(Y'Y - \hat{\beta}_A'X'X \hat{\beta}_A) = n^{-1} V(X, Y). \]

Let \( S = \sum_{i=1}^{n} Z_i Z_i' \), so that

(7) \[ S_{11} = Y'Y, \quad S_{21} = X'Y, \quad S_{22} = X'X, \]

where \( S \) is partitioned in the manner that \( \sum \) is partitioned in (1). Note \( S \) has the Wishart distribution with mean \( n \Sigma \), dimension \( p + 1 \) and \( n \) degrees of freedom. Denote the law of such a Wishart random variable by \( W(\cdot, p + 1, n) \). The ML estimator of \( \Sigma \) is \( n^{-1} S \), so that under Model B, the ML estimator of \( \Sigma^{-1} \Sigma \) is given by

(8) \[ \hat{\beta}_B = S^{-1} S_{21} = B(X, Y) \]

and the ML estimator \( \hat{\sigma}^2_B \) of \( \Sigma^{-1} \Sigma \) is given by

(9) \[ \hat{\sigma}^2_B = n^{-1}(S_{11} - S_{12} S^{-1}_{22} S_{21}) = n^{-1} V(X, Y). \]

The proof of the following theorem now follows directly from (5), (6), (8) and (9).

**Theorem 1.** The ML estimates obtained from \( \hat{\beta}_A \) and \( \hat{\sigma}^2_A \) are identical to those obtained from \( \hat{\beta}_B \) and \( \hat{\sigma}^2_B \).

This Theorem formalizes the statement that independent of the
Model, we obtain the same estimates for the regression coefficients and the variance of the error.

Note. For notational ease required in hypothesis testing, we denote the estimates corresponding to the estimators \( \hat{\beta}_A, \hat{\sigma}_A^2, \hat{\beta}_B, \hat{\sigma}_B^2 \) by \( \hat{\beta}^*, \hat{\sigma}_A^*, \hat{\beta}^*_B, \hat{\sigma}_B^* \), respectively.

The above motivation of Theorem 1 is rather algebraic in nature. A more intuitive argument can be made using conditional distributions (or likelihoods).

Note that

\[
(10) \quad f_{Y|X}(u; V; \Sigma) = f_{Y|X}(u|V; \Sigma)f_{X}(V; \Sigma) = n(u; V\beta, \sigma^2 I)f_{X}(V; \Sigma),
\]

where \( f_\cdot(\cdot; \cdot) \) denotes the density of the random variable of the subscript and the argument after the semi-colon indicates on which parameters the density depends. To calculate the ML estimates under Model B, it is necessary to maximize \( f_{Y|X}(y, X; \Sigma) \) with respect to \( \beta, \sigma^2 \) (and, peripherally, \( \Sigma \)). By the above conditioning argument, it is clear that \( f_{Y|X}(y, X; \Sigma) \) depends on \( \beta, \sigma^2 \) only through \( n(y; X\beta, \sigma^2 I) \). But the maximum of \( n(y; X\beta, \sigma^2 I) \) with respect to \( \beta, \sigma^2 \) occurs at the ML estimates of \( \beta, \sigma^2 \) under Model A.


As is to be expected, the distribution theory for \( \hat{\beta}_A \) and \( \hat{\sigma}_A^2 \) is
the same as the conditional distribution theory of \( \hat{\beta}_B \) and \( \hat{\sigma}_B^2 \) given \( X'\tilde{X} = X'X \). More interestingly, it is shown that from the relatively simpler regression analysis distribution results, we can obtain the more involved multivariate analysis of regression distribution results.

The standard distribution theory results for regression analysis (e.g., p. 333, Mood and Graybill (1963)) are

\begin{align}
11) & \quad J(\hat{\beta}_A) = N(\beta, \sigma^2(X'X)^{-1}), \\
12) & \quad J(n \hat{\sigma}_A^2/\sigma^2) = \chi^2_{n-p}
\end{align}

and

\begin{align}
13) & \quad \hat{\beta}_A \text{ and } \hat{\sigma}_A^2 \text{ are independent.}
\end{align}

In the multivariate analysis of regression case, the distribution theory results follow from the slightly more general theorem below.

**Theorem 2.** Suppose \( \mathcal{L}(S) = W(\sum, q, n) \) and that \( S \) and \( \sum \) are partitioned as in (1), with, however, the dimensions of \( \sum_{21} \) being \((q-r) \times r\). Then the following hold, where \( \sum_{11} = \sum_{11} - \sum_{12} \sum_{22}^{-1} \sum_{21} \)

\begin{align}
(1) & \quad \mathcal{L}(S_{11} - S_{12} S_{22}^{-1} S_{21}) = W(\sum_{11}, q, n-(q-r)) \\
(ii) & \quad \mathcal{L}(S_{22}) = W(\sum_{22}, q-r, n), \\
(iii) & \quad \mathcal{L}(S_{22}^{-1} S_{21} | S_{22} = s_{22}) = N(\sum_{22}^{-1} \sum_{21}, \sum_{11}^{-1} s_{22}^{-1}).
\end{align}
(where $A \otimes B$ is the Kronecker product of matrices $A$ and $B$)

(iv) $S_{11} - S_{12} S_{22}^{-1} S_{21}$ is independent of $S_{22}^{-1} S_{21}, S_{22}$.

For a proof of this theorem, see Stein (1969) pp. 29-32. This very appealing and useful theorem was first obtained by Bartlett who himself credits Wishart as the stimulus for this work. Much later proofs were given by Dempster, Stein, and Sylvan (1969). This result was also used implicitly in Anderson's (1958) derivation of the form of the Wishart distribution.

It follows from the respective definitions of the estimators under Models A and B that

$$\mathcal{L}(\hat{\beta}_A, \hat{\sigma}_A^2) = \mathcal{L}(\hat{\beta}_B, \hat{\sigma}_B^2 | \tilde{X} = X).$$

But by (11), (12) and (13), $\mathcal{L}(\hat{\beta}_A, \hat{\sigma}_A^2)$ depends on $X$ only through $X'X$, so that

$$\mathcal{L}(\hat{\beta}_B, \hat{\sigma}_B^2 | \tilde{X} = X) = \mathcal{L}(\hat{\beta}_B, \hat{\sigma}_B^2 | X'X = X'X).$$

Thus, the joint distribution of the ML estimators of the parameters under Model A is just the conditional distribution of the ML estimators under Model B given that $S_{22} = s_{22}$, where

$$s_{22} = X'X.$$

**Theorem 3.** $\mathcal{L}(\hat{\beta}_A, \hat{\sigma}_A^2) = \mathcal{L}(\hat{\beta}_B, \hat{\sigma}_B^2 | S_{22} = s_{22}).$

In deriving Theorem 2 from (11), (12) and (13) using the conditional equivalence result of Theorem 3, we make use of the following well-known proposition whose proof is trivial.
Proposition 1. If \( U \) and \( V \) are jointly distributed random vectors such that \( \mathcal{A}(U|V = v) = F(u) \), then \( U \) and \( V \) are independent and \( \mathcal{X}(U) = F(u) \).

Theorem 4. Results (11), (12), (13) imply that (i), (iii) and (iv) of Theorem 2 hold for \( r = 1 \).

Proof. With no loss of generality, let \( q = p + 1 \).

To show (i), note that from Theorem 3 and (12), \( \mathcal{L}(\hat{\sigma}^2_A, \hat{\sigma}^2_B | S_{22} = s_{22}) \)
\begin{align*}
= \mathcal{L}(n\hat{\sigma}^2_A) = \int_{11 \cdot 2} \chi^2_{n-p} , \text{ so that by Proposition 1, } \mathcal{L}(\hat{\sigma}^2_B) = \int_{11 \cdot 2} \chi^2_{n-p}.
\end{align*}

But \( \mathcal{L}(S_{11} - S_{12} S_{22}^{-1} S_{21} \mid S_{22}) = \mathcal{L}(\hat{\sigma}^2_B) \) and \( W(\int_{11 \cdot 2} 1, n-p) = \int_{11 \cdot 2} \chi^2_{n-p} \).

To demonstrate (iv), observe that

\begin{align*}
(14) \quad & \mathcal{L}(S_{11} - S_{12} S_{22}^{-1} S_{21} \mid S_{22} = s_{22}) \\
& = \mathcal{L}(\hat{\sigma}_A^2, \hat{\sigma}_B^2) I_T(s_{22}),
\end{align*}

where \( I_T(s_{22}) \) is 1 if \( s_{22} \) is in the interval, \( T \), that the law statement prescribes for \( S_{22} \) and is 0 otherwise. By (13), Theorem 3 and (i) above,

\begin{align*}
(15) \quad & \mathcal{L}(\hat{\sigma}_A^2, \hat{\sigma}_B^2) I_T(s_{22}) \\
& = \mathcal{L}(\hat{\sigma}_A^2) \mathcal{L}(\hat{\sigma}_B^2) I_T(s_{22}) \\
& = \mathcal{L}(\hat{\sigma}_B^2) \mathcal{L}(\hat{\sigma}_B^2 | S_{22} = s_{22}) I_T(s_{22}) \\
& = \mathcal{L}(S_{11} - S_{12} S_{22}^{-1} S_{21} \mid S_{22} = s_{22}),
\end{align*}

where
Now integrate the right hand side of (14) and the left hand side of (15) (which are equal) with respect to the distribution of $S_{22}$ to obtain the unconditional independence.

Finally (iii) follows directly from (11) and Theorem 3.

Note. The proof of part (ii) does not depend on parts (i), (ii) and (iv) and follows directly from the definition of the Wishart distribution and the fact that marginals of normal distributions are normal.

Theorem 2 yields that $\mathcal{L}(\hat{\beta}_B \mid S_{22} = s_{22}) = N(\beta, \sum_{11}^{-1} s_{22})$, where $\mathcal{L}(S_{22}) = W(\sum_{22}, p, n)$. Hence using one representation of the multivariate t-distribution (e.g., Representation B of Lin (1972)), we have unconditionally for $n-p > 1$,

\begin{equation}
(16) \quad \mathcal{L} \left( \frac{(n-p+1)\sum_{22}^{-1}}{\sum_{11}^{-1}} (\hat{\beta}_B - \beta) \right)^{1/2} = T_{n-p+1} (0, \sum_{22}^{-1}, p),
\end{equation}

where by $\mathcal{L}(U) = T_{\nu}(\mu, \Psi, p)$, $U, \mu: p \times 1$, $\Psi$ positive definite, $\nu > 2$, it is meant $U$ has pdf given by

\[ f(u) = C_{p, \nu} |\Psi|^{-1/2}[1 + \nu^{-1}(u - \mu)^{-1}(u - \mu)], -\nu(p)/2 \]

with $C_{p, \nu} = \Gamma[(\nu+p)/2] (\pi\nu)^{-p/2} \Gamma(\nu/2)^{-1}$.

We can use the result that if $\mathcal{L}(U) = T_{\nu}(\mu, \Psi, p)$ then
EU = μ, Cov(U) = ν (ν - 2)^{-1}ν, to obtain unconditionally

(17) \[ \mathbb{E} \hat{\beta}_B = \beta \]

(18) \[ \text{Cov}(\hat{\beta}_B) = \left[ \sum_{11 \cdot 2} / (n - p - 1) \right]^{-1} \frac{1}{22} \]

Of course, (17) and (18) can be obtained directly from Theorem 2 (and, therefore, from (11), (12), (13)). To show this, use the standard relationships between conditional and unconditional expectations, namely

\[ \mathbb{E}(\hat{\beta}_B) = \mathbb{E}(\mathbb{E}(\hat{\beta}_B \mid S_{22})) \]

and

\[ \text{Cov}(\hat{\beta}_B) = \mathbb{E}(\text{Cov}(\hat{\beta}_B \mid S_{22})) + \mathbb{E}(\text{Cov}(\hat{\beta}_B \mid S_{22})) \]

and the fact that if \( \mathcal{L}(W) = W(\gamma, p, n) \) then \( \mathbb{E}W^{-1} = (n - p - 1)^{-1} \gamma^{-1} \).

Usually \( \gamma^{-1} \) is not known, so we estimate \( \text{Cov}(\hat{\beta}_B) \) by replacing \( \gamma^{-1} \) in (18) by its unbiased estimator \( (n - p - 1) S_{22}^{-1} \). The estimated covariance is \( \sum_{11 \cdot 2} S_{22}^{-1} \) which is just the population covariance matrix of \( \hat{\beta}_A \).

In summary, it has been verified that the ML estimators of \( \sigma^2 \) have the same distribution under both Models and that the ML estimators of \( \hat{\sigma}^2 \) and \( \beta \) are independent of each other, in both cases. However, the distributions of \( \hat{\beta}_A \) and \( \hat{\beta}_B \) are different, though conditionally related. Yet under both Models, \( \mathbb{E} \hat{\beta}_A = \mathbb{E} \hat{\beta}_B \).
It is of interest to observe that all of the distributional results given in this section can be readily extended to the case when there is more than one dependent variable. The results equivalent to (11), (12) and (13) are contained in Theorem 8.2.3 of Anderson; Theorems 2 and 3 remain the same; the proof of Theorem 4 for $r > 1$ is analogous to the proof for $r = 1$; and the distribution in (16) is replaced by the matricvariate $t$-distribution. The unconditional distribution of the ML estimators of the regression coefficients for many dependent variables was first obtained by Kshirsagar (1960).

5. Hypothesis Testing: $\beta = 0$.

The question remains as to how the tests of hypotheses compare under both models. It is shown that the rejection regions and the test statistics evaluated at the data are the same for the two Models. But the two tests are shown to differ in the power function. We first consider testing $\beta = 0$ and then consider in section 6 the case of general linear constraints on $\beta$. In comparing these tests, we are led into a discussion of the relevancy to our Models of the principle of conditionality and conditional power.

We wish to test the null hypothesis that all the regression
coefficients are zero versus the alternative that they are completely general, i.e.,

(19) \( H: \beta = 0 \) (equivalently \( \sum_{22}^{1} \hat{\beta}_{21} = 0 \) )

against

\( A: \beta \neq 0 \).

Note in the multivariate analysis of regression framework, \( H \) is equivalent to testing \( \sum_{21} = 0 \), that is \( Y \) is independent of \( X \).

The standard regression theory test of (19) can be derived from the usual analysis of variance considerations. The \( \alpha \)-level test equivalent to the likelihood ratio test is to reject when

(20) \[
||x_{-A}^\hat{\beta}_A^*||^2 / ||y - x_{-A}^\hat{\beta}_A^*||^2 \geq (p/(n-p))F_{\alpha}^{1-a} \]

For convenience, write

\[
||x_{-A}^\hat{\beta}_A^*||^2 / ||y - x_{-A}^\hat{\beta}_A^*||^2 = \Phi(\hat{\beta}_A^*, \sigma_A^2, X^\top X),
\]

where

(21) \( \Phi(b, u, A) = b^\top Ab/u \),

\( b \in V_p, u \in R^+ \) and \( A \in M_p^+ \), where \( R^+ = (0, \infty) \) and \( M_p^+ \) is the set of all positive definite \( p \times p \) matrices.

The power function, \( P_A(\beta, \sigma^2) \), of this test at \( \beta, \sigma^2 \) is given by

(22) \( P_A(\beta, \sigma^2) = \text{Prob}(f_{p, n-p}^\top(\beta) \geq F_{\alpha}^{1-a}) \)
where \( L_{p,n-p}(\delta) = F_{p,n-p}(\delta) \) which is the law of the non-central F distribution with non-centrality parameter \( \delta = \frac{\beta'X'X\beta}{\sigma^2} \).

The approach followed in multivariate analysis of regression is to test for \( R^2/(1 - R^2) = \sum_{12}^{p} \sum_{22}^{p-1} \sum_{21}^{p} X'X \beta / \sum_{11}^{p} \beta \beta / \sum_{11}^{p} \sigma^2 \) being zero. \( R \) is the multiple regression coefficient of \( Z_1 \) on \( Z_2, \ldots, Z_{p+1} \) and is defined by \( R = (\sum_{12}^{p} \sum_{22}^{p-1} \sum_{21}^{p} Y'Y)^{1/2} / (\text{Var } Z_1)^{1/2} \),

where \( Z_1 \) is the \( i \)th entry of \( Z \). The statistic employed here is \( \hat{R}^2/(1 - \hat{R}^2) \) which is equal to \( \phi(\hat{\beta}_B, \hat{\sigma}_B^2, S_{22}) \), \( \phi \) being defined by (21). One representation of the distribution of \( \hat{R}^2/(1 - \hat{R}^2) \) is given by Stein as

\[
L(\hat{R}^2/(1 - \hat{R}^2)) = \chi^2_{p+2K} / \chi^2_{n-p} ;
\]

where \( K \) has the negative binomial distribution

\[
P(K = k) = \frac{\Gamma(k + n/2)\Gamma(k)^{-1}\Gamma(n/2)^{-1}\Gamma(\hat{R}^2)^{-k/2}\Gamma(1 - \hat{R}^2)^{-n/2}}{\Gamma(n/2)^{-1}} \]

and \( \chi^2_{p+2K} \) is independent of \( \chi^2_{n-p} \). Under \( H, R = 0 \) which implies \( K = 0 \) (almost surely), so that an \( \alpha \)-level test of (19) is to reject if

\[
\phi(\hat{\beta}_B, \hat{\sigma}_B^2, S_{22}) \geq (p/(n-p))F_{p,n-p}^{1-\alpha} \]

Of course the test is again equivalent to the likelihood ratio test. For this test the power function, \( P_B(\beta, \sigma^2, \sum_{22}) \), is given by
\[ (25) \quad P_B(\beta, \sigma^2, \sum_{22}) = \text{Prob}(\xi \geq (p/(n-p))F_{p,n-p}^{1-\alpha}) \]

where \( \mathcal{L}(\xi) \) is given by (23).

From (23) it can be shown that

\[ \mathcal{L}(R^2/(1-R)^2|S_{22} = s_{22}) = (p/(n-p))F_{p,n-p}(\delta), \]

where

\[ \delta = \sum_{12} \sum_{22}^{-1} s_{22} \sum_{22}^{-1} \sum_{21} / (\sum_{11} - \sum_{12} \sum_{22}^{-1} \sum_{21}) \]

\[ = \beta^- s_{22} \beta^+ / \sum_{11} \cdot 2 \]

Thus, \( P_A(\beta, \sigma^2) \) is the power of the conditional test (24) given \( S_{22} = s_{22} \), i.e.,

\[ P_A(\beta, \sigma^2) = \text{Prob}(\xi \geq (p/(n-p))F_{p,n-p}^{1-\alpha}|S_{22} = s_{22}) \]

From Theorem 1 and the above, we obtain the following theorem.

**Theorem 5.** The \( \alpha \)-level tests of \( H_0: \beta = 0 \) against \( H_a: \beta \neq 0 \) given by (20) and (24) have identical critical functions and have power functions given by (22) and (25), respectively.

**Corollary 1.** The \( 100(1-\alpha)\% \) simultaneous confidence regions corresponding to the tests (20) and (24) are identical and given by (using Model B notation)
\[ n^{-1}(\hat{\beta}^* - \beta)^\top \hat{s}_{22}(\hat{\beta}^* - \beta) / \hat{s}_B^2 \leq (p/(n-p))F_{p,n-p}^{1-\alpha} \]

As before, one can derive results for the unconditional model from the conditional model. In this case one could verify the implication using the terminology of conditional likelihoods as is done in the end of section 3. A more appropriate setting for the terminology lies in the concept of the principle of conditionality.

We give a short description of what we mean by this principle. This definition is found in Stein p. 43.

Let \( E_c, c \in C \), be a set of experiments and let \( Q \) be a probability measure on \( C \). Suppose for \( c \in C \), we observe \( X_c \) with a distribution \( P_\theta \) such that when \( \theta = \theta_0 \), \( P_{\theta_0} \) does not depend on \( c \in C \). Suppose also that an appropriate level \( \alpha \) test for \( H_0: \theta = \theta_0 \) based on \( E_c \) is to reject if \( X_c \in S_c \).

Let \( C \) be a random variable distributed according to \( Q \) in \( C \). Suppose that we choose \( C \) according to \( Q \) and then perform \( E_C \) to obtain \( X_C \). The principle of conditionality asserts that an appropriate level \( \alpha \) test is to reject \( H_0 \) if \( X_C \in S_C \).

In Model B, we can view \( E_C \) as that experiment which results in \( S_{22} \). Then associate \( \Phi_A \equiv \Phi(\hat{\beta}, \hat{s}_A^2, X^\top X) \) with \( X_C \). Under \( H: \beta = 0 \), \( \Phi_A \) has a distribution which does not depend on \( c \in C \). Thus, the principle of conditionality says the appropriate \( \alpha \)-level test of \( H \) is given by (24).
However, the principle of conditionality asserts nothing about how to deal with power. The question remains whether or not to deal with

\[(26) \quad \text{Prob}(X_c \in S_c; \theta)\]

or

\[(27) \quad \text{Prob}(X_c \in S_C; \theta),\]

respectively, the conditional power (i.e., the power of the conditional test) and the unconditional power. Birnbaum (1962) states that we should deal with the conditional power as long as \(\mathcal{L}(C)\) does not depend on \(\theta\).

Lehmann (1959, pp. 139-140) discusses an interpretation for viewing (26) as the appropriate power but points out that "the disadvantage [of (26)] compared with an unconditional power [is] that it becomes available only after the observations have been taken. It therefore cannot be used to plan the experiment and in particular to determine the sample size, if this must be done prior to the experiment." Thus, in our regression situation, if one is interested in planning experiments utilizing power, it is necessary to deal with \(P_B\) instead of \(P_A\). However, Lehmann does suggest that if \(\mathcal{L}(C)\) depends on some unknown parameter different from \(\theta\), then (26) can be viewed as an estimate of (27). Such a viewpoint corresponds to the
discussion in section 4 of estimating the unconditional law of \( \hat{\beta}_B \) by the conditional law of \( \hat{\beta}_{A} \).

To return, let us note that (25) can be obtained from (22) by integrating (22) against the Wishart density of \( S_{22} \). Such an integration is done in a more general setting in section 6. We now have the following.

**Theorem 6.** Given that (20) is a test of (19) under Model A with power function given by (22), then under Model B, (24) is a test of (19) with power function given by (25).


The technique for testing hypotheses about general linear constraints on \( \beta \) is well known under Model A. Theorem 7 which is given in Rao (1965, pp. 155-6) presents, among others, the major results necessary to derive such a test of hypothesis and its power function. However under Model B, the corresponding test is usually not discussed in standard texts. We employ Theorems 2 and 3 to derive a test of general linear constraints for Model B from the results of Model A. Particular attention is paid to deriving the unconditional power function.

**Theorem 7.** Suppose \( H \in M_{p,k} \), \( k \leq p \), is such that the column space generated by \( H \) is a subspace of the column space generated by \( X'X \) (which in this paper is assumed to be of full rank).
Let

$$L_0^2 = \min_{\beta} \|Y - X\beta\|^2$$

and

$$L_1^2 = \min_{\beta} \|Y - X\beta\|^2,$$

where this minimization is subject to the constraint $H'\beta = \xi$ (given). Then

(a) $L_0^2$ and $L_1^2 - L_0^2$ are independently distributed,

(b) $\mathcal{A}(L_0^2) = \sigma^2 \chi^2_{n-p}$,

(c) $\mathcal{A}(L_1^2 - L_0^2) = \sigma^2 \chi^2_k(\delta)$, where

$$\delta = (H'\beta - \xi)'(H'\beta - \xi) / \sigma^2.$$

(d) If $H'\beta = \xi$ is true, then

$$\mathcal{A}(L_1^2 - L_0^2)/L_0^2 = (k/(n-p))F_{k,n-p}.$$

Note that under our assumption that $X'X$ is nonsingular, $L_0^2 = n \sigma^2$.

Using Theorems 2 and 3, we can now obtain under Model B, a test for

$$(28) \quad H : H' \sum^{-1}_{22} \sum_{21} = \xi$$

against

$$A : H' \sum^{-1}_{22} \sum_{21} \neq \xi,$$

where $H$ is $p \times k$, $p \geq k$, and of full rank.
Let \( \hat{\lambda}_0^2 = n \hat{\sigma}_B^2 = S_{11} - S_{12} S_{22}^{-1} S_{21} \) and \( \lambda_1^2 = \min_{H' \beta = \bar{x}} \| Y - X\beta \|^2. \)

Further define \( \lambda_0^{2*} = n \hat{\sigma}_B^{2*} \) and \( \lambda_1^{2*} = \min_{H' \beta = \bar{x}} \| y - X\beta \|^2. \)

**Theorem 8.** An \( \alpha \)-level test of the hypothesis of (28) is to reject if

\[
(\lambda_1^{2*} - \lambda_0^{2*})/\lambda_0^{2*} \geq (k/(n - p))F_{k,n-p}^{1-\alpha}.
\]

**Proof.** Assume that \( H' \beta = \bar{x} \) is true. Then by Theorem 3 and (b) of Theorem 7, \( \mathcal{L}(\lambda_0^2 | s_{22} = s_{22}) = \sum_{11.2} \chi^2_{n-p} \). Noting that the law of \( \lambda_1^2 - \lambda_0^2 \) depends on \( X \) only through \( X'X \), we have, by Theorems 3 and 7 (c), (d) that \( \mathcal{L}(\lambda_1^2 - \lambda_0^2 | s_{22} = s_{22}) = \sum_{11.2} \chi^2_{k} \).

Hence, by Proposition 1, unconditionally

\[
(29) \quad \mathcal{L}(\lambda_1^2 - \lambda_0^2) = \sum_{11.2} \chi^2_{k} \text{ and } \mathcal{L}(\lambda_0^2) = \sum_{11.2} \chi^2_{n-p}.
\]

To complete the proof we need to show that \( \lambda_0^2 \) and \( \lambda_1^2 - \lambda_0^2 \) are unconditionally independent for all possible parameter values.

From Theorem 7 (a), we have

\[
(30) \quad \mathcal{L}(\lambda_1^2 - \lambda_0^2 | s_{22} = s_{22}) = \mathcal{L}(\lambda_1^2 | s_{22} = s_{22}) \mathcal{L}(\lambda_0^2 | s_{22} = s_{22}) = \mathcal{L}(\lambda_1^2 | s_{22} = s_{22}) \mathcal{L}(\lambda_0^2).
\]
this last equality following directly from Theorem 2 (iv) or again from Theorem 7 (b) and Proposition 1. Integrate the first and last expressions of (30) against $S_{22}$ to yield the desired independence result.

Note that had we not wanted to establish the results corresponding to Theorem 7 (a), (b), (d) for Model B, the proof of Theorem 8 follows immediately from the principle of conditionality and Theorem 7 (d).

As is expected, the critical function of the test of (26) under Model B is the same as that under Model A. Again it is in the power functions that the differences occur. To this end, we state certain elementary results concerning the inverse Wishart distribution, so that the distribution, under the alternative $H \sum_{21}^{2} \lambda_{0}^{2} \neq \xi$, of $(\lambda_{1}^{2} - \lambda_{0}^{2})/\lambda_{1}^{2}$ can be found.

**Definition.** $\psi$ has the inverse Wishart distribution, denoted $W^{-1}(\Lambda, p, n)$, if and only if $\mathcal{L}(\psi^{-1}) = W(\Lambda^{-1}, p, n)$.

**Lemma 1.** If $\mathcal{L}(\psi) = W^{-1}(\Lambda, p, n)$, then for $A \in M_{P,P}^r$, $\mathcal{L}(A \psi A^{-1}) = W^{-1}(A \Lambda A^{-1}, p, n)$.

**Proof.** $\mathcal{L}((A \psi A^{-1})^{-1}) = \mathcal{L}(A^{-1} \psi^{-1} A^{-1}) = W((A \Lambda A^{-1})^{-1}, p, n)$.

**Lemma 2.** Suppose $\psi$ and $\Lambda$ are partitioned as in (1) with, however, $\psi_{21}$ being $(p - q) \times q$. If $\mathcal{L}(\psi) = W^{-1}(\Lambda, p, n)$, then
\[ \mathcal{L}(\psi_{11}) = W^{-1}(\Lambda_{11}, q, n - (p - q)). \]

**Proof.** Let \( \psi_{11} \) be partitioned in the same manner as \( \psi \). Then \( \psi_{11}^{-1} = W_{11} W_{12}^{-1} W_{22} W_{21}^{-1} \), so that by Theorem 2 (i),

\[ \mathcal{L}(\psi_{11}^{-1}) = W(\sum_{11} - \sum_{12} \sum_{21}^{-1} \sum_{22}, q, n - (p - q)). \]

Thus

\[ \mathcal{L}(\psi_{11}) = W^{-1}(\Lambda_{11}, q, n - (p - q)), \]

because

\[ \Lambda_{11} = (\sum_{11} - \sum_{12} \sum_{21}^{-1} \sum_{22})^{-1}, \]

for \( \Lambda = \sum^{-1} \).

Lemmas 1 and 2 now provide the tools to obtain the unconditional distribution of \( \lambda_{11}^2 - \lambda_0^2 \). The next two lemmas enable the finding of the distribution of \( \tau^2(S_{22}) \), where \( \mathcal{S}(S_{22}) = W(\sum_{22}, p, n) \) and we define \( A \in M_{p,p}^+ \)

\[ \tau^2(A) = (H^*H - \xi)(H^*A^{-1}H)^{-1}(H^*H - \xi)/\sum_{11} \mathcal{L}. \]

**Lemma 3.** \( \mathcal{L}(H^*S_{22}^{-1} H) = W((H^*H)^{-1}, k, n - (p - k)). \)

**Proof.** Since \( H \) is of full rank, there exists a \( p \times (k - p) \) matrix \( L \) such that \( H^* = (H; L) \) is nonsingular. By Lemma 1,

\[ \mathcal{L}(H^*S_{22}^{-1} H) = W^1(H^*H^*L^{-1} H*, p, n), \]

so that by Lemma 2,

\[ \mathcal{L}(H^*S_{22}^{-1} H) = W^{-1}(H^*H^*L^{-1} H, k, n - (p - k)), \]

yielding the result.

**Lemma 4.** \( \mathcal{L}(\tau^2(S_{22})) = \tau^2(\sum_{22}) \chi^2_{n - p + k} \).
Proof. This result follows directly from the fact that if $\mathcal{L}(S) = W(\lambda, p, n)$ then $\mathcal{L}(\lambda - \lambda^2) = (\lambda - \lambda^2) \chi^2_n$.

Theorem 2. $\mathcal{L}(\lambda_1^2 - \lambda_0^2) = \sum_{1 \leq 2} \chi^2_{k+2n}$, where $n$ is a random variable independent of $\lambda^2$ and for $i = 0, 1, \ldots$,

$$\text{Prob}(\eta = i) = \frac{\Gamma(\frac{n-p+k}{2} + i)}{\Gamma(i+1) \Gamma(\frac{n-p+k}{2})} \left(1 + \tau^2(\sum_{22})\right)^{-\frac{(n-p+k)/2}{2}} \left(\frac{\tau^2(\sum_{22})^i}{1 + \tau^2(\sum_{22})}\right).$$

Proof. Noting that $\mathcal{L}(\chi^2_k(\delta)) = \mathcal{L}(\chi^2_{k+2\delta})$, where $\mathcal{L}(\delta)$ is Poisson with parameter $\delta/2$, we have from Theorem 7 (c) that

$$\mathcal{L}(\lambda_1^2 - \lambda_0^2)/\sum_{1 \leq 2} |S_{22} = s_{22}| = \mathcal{L}(\chi^2_k(\tau^2(s_{22})))|\tau^2(s_{22}) = \tau^2(s_{22})) = \mathcal{L}(\chi^2_{k+2n}|\tau^2(s_{22}) = \tau^2(s_{22})),$$

where $\mathcal{L}(\eta|\tau^2(s_{22}) = \tau^2(s_{22}))$ is Poisson with parameter $\tau^2(s_{22})$.

Thus, $\mathcal{L}(\lambda_1^2 - \lambda_0^2)/\sum_{1 \leq 2}$ is $\mathcal{L}(\chi^2_{k+2n})$, where by Lemma 4, the unconditional distribution of $\eta$ is given by

$$\text{Prob}(\eta = i) = \int_0^\infty \frac{x^{(n-p+k)/2-1}}{\Gamma(i+1) \Gamma(\frac{n-p+k}{2})} \left(\frac{x}{\tau^2(\sum_{22})}\right)^{(n-p+k)/2} \frac{-x/\tau^2(\sum_{22})}{\Gamma(\frac{n-p+k}{2})},$$

which is the distribution of (3).


Corollary 2. \( \mathcal{A}(\lambda_1^2 - \lambda_0^2) = \chi^2_{k+2n}/\chi^2_{n-p} \), where \( \mathcal{A}(\cdot) \) is given by (22) and \( \chi^2_{k+2n} \) is independent of \( \chi^2_{n-p} \).

Proof. This corollary follows directly from Theorem 9, (29) and (30).

The power function of the test described in Theorem 8 is
\[
P^*_B(\theta, \sum_{22}, \sum_{11}^2) = \text{Prob}(\zeta \geq (k/(n-p))F_{k,n-p}^{1-\alpha}),
\]
where \( \mathcal{A}(\cdot) = \chi^2_{k+2n}/\chi^2_{n-p} \), where the numerator and the denominator are independent and \( \mathcal{A}(\cdot) \) is given by (31).

7. Remarks.

There are other derivations of the type pursued in section 6, where "new" results can be obtained. For example, one test of hypothesis that is often considered is to test equality of regression coefficients (e.g., see Rao (1965) pp. 237-239 or Fisher (1970)). It is possible to view the test described by Rao as being conditional on \( S_{22} \) and \( T_{22} \), these being the covariance matrices of the independent variables for each sample. To obtain the equivalent test under assumptions akin to Model B, we would find the power function by integrating against the distributions of \( S_{22} \) and \( T_{22} \).

For this testing problem, Theorem 2 (iii) can be used to provide an answer. We have
\[
\mathcal{A}((\hat{\beta}_1 - \hat{\beta}_2 | s_{22} = s_{22}, T_{22} = t_{22}) = N(\beta_1 - \beta_2, \sigma^2(s_{22}^{-1} + t_{22}^{-1}))
\]
where it is assumed that $\sum_{11.2}^{(s)} = \sum_{11.2}^{(t)} = \sigma^2$. Then we invoke

the standard normal test of $\mu = 0$ which would now be conditional

on $T_{22}$ and $S_{22}$. Integration against $S_{22}$ and $T_{22}$ provides the power

function for this test.

This paper was solely concerned with one dependent variable.
As is noted in section 4 an extension to the case where there are
many dependent variables (each correlated with the other) is possible.
In this multi-dependent variable situation, the model analogous to
Model A would be the corresponding multivariate analysis of variance
formulation (e.g., see Anderson pp. 178-229). The model analogous
to Model B would be exactly the extension of the $Z_1$ of Model B to $p$
independent variables and $q$ dependent variables being jointly nor-

mally distributed. For this multi-dependent variable regression
problem, there are $U$ statistics corresponding to $F$ statistics and
to obtain the unconditional version of the $U$ statistic for the new
Model B version, the usual integration is necessary. Of course it
would be the case in the multi-dependent variable regression models,
that the statistical analyses are conditionally related in the same
way the analyses considered here are related.
References


A Tale of Two Regressions

Technical Report: July 1973

Allan R. Sampson

July 1973

N000 14-67-A-0235-0006

OMR Technical Report No. 75

PSU Statistics Report M263

Releasable without limitations on dissemination.

This expository paper details the effects of the choice of underlying model in the statistical analysis of regression. The two models considered are the cases when all the data is jointly normally distributed and when the independent variables are assumed to be constants. The standard results for both these models are related to each other conditionally in a unified manner and an example is given where new results in one model can be obtained from old results in the other model. Particular emphasis is devoted to demonstrating differences that arise in power function considerations.
Regression
Multivariate normal distribution
Power
Conditional tests of hypotheses
multivariate t-distribution