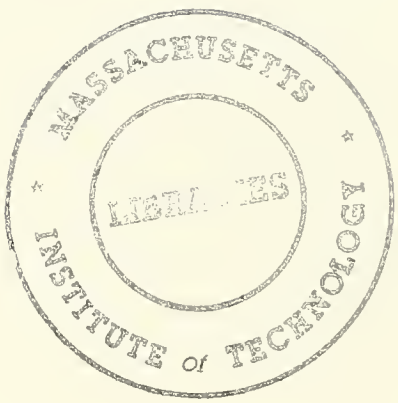

MASSACHUSETTS
INSTITUTE of TECHNOLOGY

working paper
department
of economics

An Exact Small Sample Test of Non-Nested Hypotheses

by

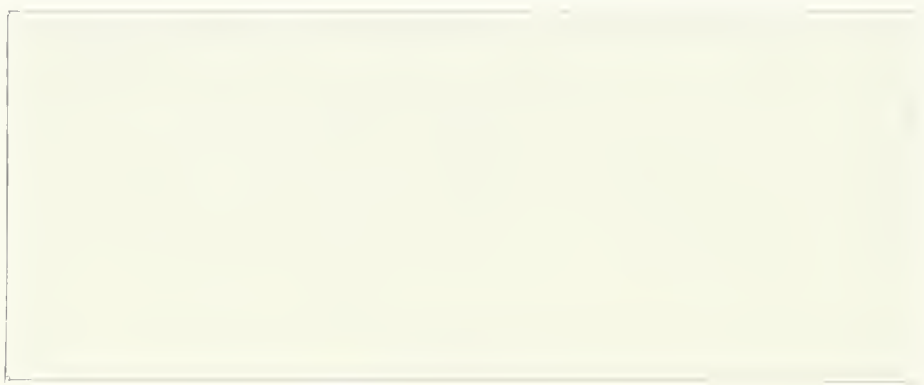
Robert Feenstra
Martin L. Weitzman

No. 275

December, 1980

massachusetts
institute of
technology

50 memorial drive
cambridge, mass. 02139



An Exact Small Sample Test of Non-Nested Hypotheses

by

Robert Feenstra
Martin L. Weitzman

No. 275 December, 1980

An Exact Small Sample Test of Non-Nested Hypotheses

Summary

Classical hypothesis testing typically validates or invalidates a model by nesting it in a more general model and performing a likelihood ratio test. When models are not nested there does not seem to exist a practical way of comparing them. Generalizing the classical statistical methodology of the likelihood ratio, we develop a practical, exact, small sample test of non-nested hypotheses. The test may be useful because it can allow one to check the validity of a specification directly. As an application, we find that time series data may not support the existence of an aggregate production function for the U.S. economy.

Introduction

All of us have been exposed to situations where an elaborate econometric model, supposedly based on theory, hardly fits the data better than some simple minded ad hoc specification. Given any model, if there are easily available alternative ways of organizing data which yield results "too good to be true" in some sense, that should cast doubt on the authenticity of the original specification. The discrediting alternative need not be theoretically justified; it need not pretend to any theoretical content whatsoever. Intuitively, there ought to be an exact sense in which something is suspicious about a model that is supposed to be correct, when casually chosen alternatives also fit so well.

This approach both implies and is implied by a particular philosophy of modeling which can be simply stated as follows: should we find that a particular model implies unlikely consequences, it should lead us to question or even perhaps to reject the underlying specification.

The fundamental barrier to formalizing such ideas has been the lack of an operational test for non-nested hypotheses. A procedure like the t-test, for example, automatically screens out such foolish behavior as accepting a linear model with a zero coefficient when an obvious alternative fits the data much better. But the inability to compare non-nested hypotheses is a very severe limitation which prevents us from inferring that a model as a whole may be incorrectly specified when some other model is fitting too well relative to it. Nor does the trick of

artificially nesting two equally well fitting alternatives in a more general model help much in practice, because multi-collinearity will almost surely prevent either alternative from rejecting the other.

In this paper we offer an exact small sample test of non-nested models which directly generalizes the classical F-test. In most cases there should be no difficulty in performing our test routinely on the computer, although it is definitely a more demanding computation than the classical tests to which it reduces in a nested environment.

The test statistic we use is the likelihood ratio proposed by Cox [1961], [1962], and subsequently applied by Pesaran [1974]. Our own philosophy of modeling closely parallels the exposition of Pesaran and Deaton [1978], q.v., except that we would go further in asserting the "other" hypothesis can be used to reject a maintained hypothesis even without any pretence whatsoever about the viability of the rejection model as an alternative.

The main contribution of the present paper consists of showing that there is essentially no computational barrier to getting as close as we want to the exact distribution of the likelihood ratio statistic for non-nested linear models. By looking at the problem slightly unconventionally, it is possible to obtain an exact, small sample test which avoids altogether the pitfalls of asymptotic large sample theory.

Our feeling is that the test we propose is powerful and is likely to prove damaging to certain classes of models. As an example, we show that the existence of a stable aggregate production function for the U.S. economy may be questionable because a general formulation with capital

and labor explains the growth of output significantly worse (or insignificantly better) than a specification which is identical except for omitting capital and labor altogether.

Note that in our framework it does not make sense to talk about proving a model is "true". A model can never be proved true; it can only be discredited, which occurs if it is found to result in unlikely coincidences. A model remains a viable hypothesis only under circumstances in which other data or other models do not yield an implausibly good fit, considering the original model is supposed to be correct. Models are viable, as it were, only by default. Any given field may or may not be characterized by a viable model, and the situation can change over time. In our view this is a correct description of how science works. It is quite possible to be in an agnostic situation, temporarily or permanently, where no single hypothesis is able to establish itself against refutation. If a model has established itself (by default) as viable, it may at any time be discredited by newly conceived specifications or changed data. And the discrediting model does not have to be viewed as a candidate to replace the discredited model, since it too may be non-viable.

Formal Description of the Test

The model being tested, called the "hypothesis model" is denoted (H). Another model, denoted (R), may be used to possibly reject (H). It is assumed that the hypothesis and rejection models satisfy:

$$(H)\quad \underset{(T\times 1)}{Y} = \underset{(T\times k_x)}{X}\ \underset{(k_x\times 1)}{\beta} + \underset{(T\times 1)}{\varepsilon} \qquad X \text{ and } \varepsilon \text{ independent}$$

$$\varepsilon_i \text{ iid } N(0,\sigma^2),$$

$$(R)\quad \underset{(T\times 1)}{f(Y)} = \underset{(T\times k_z)}{Z}\ \underset{(k_z\times 1)}{\gamma} + \underset{(T\times 1)}{\delta} \qquad Z \text{ and } \varepsilon \text{ independent.}$$

X and Z represent data that might conincide in part, or might be completely different. f(Y) denotes some transformation of Y, such as $f_i(Y_i) = \log Y_i$, or $f(Y) = Y + V$ for some vector V. The most common transformation used in practice would be the identity $f(Y) = Y$. The rejection model (R) is not assumed to be a reasonable "alternate hypothesis" to (H), or to satisfy any statistical properties beyond independence of Z and ε.

Suppose we consider performing OLS regressions on (H) and (R), and using the ratio of the sum of squared residuals (SSR) as a test statistic λ:

$$SSR_H = (Y-X\hat{\beta})'(Y-X\hat{\beta}) = Y'M_X Y$$

$$\hat{\beta} = (X'X)^{-1}X'Y\ ,\ M_X = I - X(X'X)^{-1}X'$$

$$SSR_R = (f(Y)-Z\hat{\gamma})'(f(Y)-Z\hat{\gamma}) = f(Y)'M_Z f(Y)$$

$$\hat{\gamma} = (Z'Z)^{-1}Z'f(Y)\ ,\ M_Z = I - Z(Z'Z)^{-1}Z'$$

and

$$\lambda = \frac{SSR_R}{SSR_H} = \frac{f(Y)'M_Z f(Y)}{Y'M_X Y} \quad . \tag{2}$$

Under the assumption that (H) is the correct model, suppose we could somehow obtain the distribution $G(\lambda)$ of our test statistic. Then, following the classical likelihood-ratio approach, we could specify a critical region $[0,c]$, and reject model (H) if $\lambda \in [0,c]$. Choosing $c = \lambda$, we would always reject (H), and the probability that this decision is wrong is given by

$$\alpha = \int_0^c dG(\lambda).$$

That is, α is the probability of making a type I error, or the level of significance of the test.

In this framework, we reject (H) at a high level of significance (low α) only when the test statistic λ is much lower than we would expect if (H) were true. That is, should we find that assuming (H) true leads to an unlikely event (i.e., low λ), we conclude that the hypothesis model cannot be the correct specification. Our test is thus a formalization of what we mean by a model being "untrue", i.e., its acceptance leads to unlikely consequences.

The distribution $G(\lambda)$ is derived as follows. Assuming that (H) is true, we can substitute $Y = x\beta+\varepsilon$ into (2) to obtain,

$$\lambda = \frac{f(X\beta+\varepsilon)'M_Z f(X\beta+\varepsilon)}{\varepsilon' M_X \varepsilon} \quad . \tag{2'}$$

Now suppose that β did not appear in (2'), e.g., suppose we could rewrite the expression so that β cancels out (this actually occurs in the classical case of nested hypotheses). Then one method of obtaining the distribution of λ is to simply simulate ε using a random number generator. Denoting simulated values with a tilde we would have

$$\tilde{\lambda} = \frac{f(X\beta+\tilde{\varepsilon})'M_Z f(X\beta+\tilde{\varepsilon})}{\tilde{\varepsilon}' M_X \tilde{\varepsilon}} \tag{3}$$

Repeating this simulation many times, we would obtain a frequency distribution which has been derived under the assumption that model (H) is correct, so with enough simulations $\tilde{G}$ would approach the actual distribution G. Our hypothesis test could then be performed.

Of course, the unknown coefficients β need not cancel out of (2') or (3), except in the special case of nested models. But under the assumption that (H) is correct, $\hat{\beta}$ has a precise relation to β and ε, namely,

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta+\varepsilon) = \beta + (X'X)^{-1}X'\varepsilon$$

Using the assumption of "flat" priors on β (which underlies classical statistics), we can use Bayes' law to invert the above expression, obtaining a posterior distribution of β as a function of ε, conditional on the observed $\hat{\beta}$:

$$\beta = \hat{\beta} - (X'X)^{-1}X'\varepsilon. \tag{4}$$

In our simulations, posterior values of β are calculated from the formula $\tilde{\beta} = \hat{\beta} - (X'X)^{-1}X'\tilde{\varepsilon}$. Substituting $\tilde{\beta}$ for β in (3), we derive

$$\tilde{\lambda} = \frac{f(X\hat{\beta} + M_X\tilde{\varepsilon})'M_Z f(X\hat{\beta} + M_X\tilde{\varepsilon})}{\tilde{\varepsilon}'M_X\tilde{\varepsilon}}. \tag{3'}$$

From (3') we see that each simulation merely requires the evaluation of linear and quadratic forms in $\tilde{\varepsilon}$ with a fixed matrix and does <u>not</u> require inverting a new matrix at each step. This crucial feature makes our exact, small sample test an absolutely routine calculation. It is derived from two underlying assumptions. First, the (H) and (R) models must be linear in β and γ. Second, X and Z must be independent of ε, since otherwise when we simulate ε we must also simulate X and Z, and re-compute M_X and M_Z. Note, however, that transformations of the form f(Y) can easily be performed. Our experience is that the test can be calculated in a routine fashion on the computer and could be programmed as a standard part of a regression package.

Our method for obtaining $\tilde{G}$ is now complete except for one detail: in order to generate $\tilde{\varepsilon}$ we must know its variance. The ε_i are assumed to be iid $N(0,\sigma^2)$, and σ^2 is estimated as

$$\hat{\sigma}^2 = \frac{Y'M_X Y}{(T-k_X)}$$

Under the assumption that (H) is the correct model, we have

$$\hat{\sigma}^2 = \frac{\varepsilon' M_X \varepsilon}{(T-k_x)} , \qquad \frac{\varepsilon' M_X \varepsilon}{\sigma^2} \overset{d}{=} \chi^2(T-k_x).$$

Consistent with classical statistics, uniform or "flat" priors on $\log\sigma^2$ yields the posterior distribution,

$$\sigma^2 \overset{d}{=} \frac{(T-k_x)\hat{\sigma}^2}{\chi^2(T-k_x)} . \tag{5}$$

In practice, posterior values of σ^2 are calculated from the formula $\tilde{\sigma}^2 = (T-k_x)\hat{\sigma}^2/\tilde{\chi}^2$, where $\tilde{\chi}^2$ is obtained from a random-number generator. For each drawing of $\tilde{\chi}^2$ we generate a vector $\tilde{\varepsilon}$, and then compute $\tilde{\lambda}$ from (3'). Repeating this simulation many times we obtain the distribution $\tilde{G}$, derived under the assumption that (H) is correct. The hypothesis test described above can thus be routinely performed on the computer. We are calculating the probability α that the observed test statistic λ should be as small as it is, given the maintained hypothesis (H). The error in estimating α can easily be calculated from the binomial distribution: with N simulations, $\text{var}(\tilde{\alpha}) = \alpha(1-\alpha)/N$.

Our procedure for testing (H) using (R) shows why the rejection model itself need not be believable: at no point have we assumed that (R) satisfies any desirable statistical properties, beyond independence of Z and ε. This feature is closely related to our methodological view that models can be rejected, but not accepted (except by default). To formally accept a model, we need to know the probability that this decision is wrong (i.e., the type II error). In other words, we must know the distribution of our test statistic λ when (H) is <u>not</u> the correct specification. But if (H) is not correct, then we do not know the "true"

specification: in some cases (R) might be a reasonable candidate, but there are surely many other possibilities. In principle, one can imagine assigning priors to all possible specifications, and then integrating over the space of models to obtain the probability of a type II error. In practice, such a computation could obviously not be performed. Note that reversing the roles of (H) and (R) simply means that (H) could be used to reject, but not accept, model (R).

While we have used some elementary Bayesian methodology in deriving the distribution $G(\lambda)$, our procedure for testing non-nested hypotheses is really no more Bayesian than any other aspect of classical statistics. In fact, our test is a direct generalization of the classical, nested likelihood-ratio tests.

The classical statistical approach to analyzing the regression model $Y = X\beta+\varepsilon$ is consistent with the Baynesian approach of using a quadratic loss function, and diffuse priors on β and $\log\sigma^2$. For example, the standard t-test for the hypothesis $\beta_1 = b_1$ is equivalent to obtaining the posterior distributions (4) and (5), and then testing whether b_1 is likely to have been drawn from the distribution of β_1. That is, we exclude the upper and lower $100\alpha/2$ percentiles of β_1, and then observe whether b_1 falls into the middle range. This approach is computationally identical to the t-test. Observe, however, that our formulation does not depend on a fortuitous cancelling out of β in (2') which may make it a more appealing conceptual approach for understanding statistical

testing, even for nested hypotheses.

We note that for the special case of nested models, our procedure reduces to the standard F or t tests. As an example, consider testing the hypothesis that certain coefficients of a model equal zero. Then the data matrix X of (H) includes only a subset of the variables in Z. It follows that $M_z X = 0$, and so with $f(Y) = Y_*(2')$ becomes

$$\lambda = \frac{\varepsilon' M_z \varepsilon}{\varepsilon' M_x \varepsilon} .$$

That is, the test statistic λ is simply the SSR of the unrestricted model divided by the SSR of the restricted model. A simple transformation of λ, i.e.

$$\frac{(1-\lambda)/(k_z - k_x)}{\lambda/(T-k_z)}$$

is distributed as $F(k_z-k_x, T-k_z)$.[1]

That classical nested hypothesis testing with its well known statistical properties is a special case of our generalization suggests to us the general test is likely to be "powerful" in some sense. We expect that to pass the λ test, the hypothesis model in most cases has to perform somewhat better than a casually specified, empirically oriented rejection model.

More work is needed to determine the properties of our test under general and specific circumstances. As one special case, consider testing a model against itself (i.e., (H) = (R)). Substituting f(Y) = Y and Z = X into (3'), the simulated test statistic becomes

$$\tilde{\lambda} = \frac{\tilde{\varepsilon}' M_x \tilde{\varepsilon}}{\tilde{\varepsilon}' M_x \tilde{\varepsilon}} \equiv 1.$$

The observed test statistic is also $\lambda = 1$. In other words, we are observing an event which occurs with probability one, and so we certainly cannot reject (H) (α=1). This suggests that testing a model using another which is "similar" will not lead to a rejection. Intuitively, some sort of orthogonality between X and Z, along with a good fit for (R), should lead to stronger tests (lower α).

Finally, note that while we are allowed to choose (R) by comparing the X and Z matrices, we definitely cannot choose the rejection model by first observing the results of the test, and then searching for a Z with stronger results: the latter procedure would violate the assumption of independence between Z and ε.

An Application

Without taking the example too seriously, the aggregate production function nicely illustrates some of the points we are trying to make.

In the aggregate U.S. private economy for year t, let

$Y(t)$ = actual output

$\hat{Y}(t)$ = potential output

$E(t)$ = employment rate

$L(t)$ = labor

$K(t)$ = capital

Suppose we postulate an Okun's law type of relationship between potential and actual output of the form

$$Y(t) = \hat{Y}(t) \quad E(t)^{\gamma}. \tag{6}$$

By this rule a one point increase in unemployment decreases output by $\gamma\%$.

If potential output is accurately described by a constant returns to scale production function of the general form $F(K,L)$ with Hicks neutral technical change, then

$$\hat{Y}(t) = A(t)\ F(K(t),L(t)). \tag{7}$$

In practice we have chosen $A(t)$ to be of the exponential form

$$A(t) = A\, e^{\lambda t}, \tag{8}$$

and verified that more general specifications like

$$A(t) = A\, e^{\lambda' t + \lambda'' t^2} \tag{9}$$

do not negate our conclusions.

Logarithmically differentiating (6), (7), and (8), we can derive the "growth equation"

$$g_y = (\eta_k g_k + \eta_\ell g_\ell) + \lambda + \gamma g_e. \tag{10}$$

In formula (10),

$$g_y = \frac{\dot{Y}}{Y}$$

$$g_k = \frac{\dot{K}}{K}$$

$$g_\ell = \frac{\dot{L}}{L}$$

$$g_e = \frac{\dot{E}}{E}$$

and

$$\eta_k = \frac{K \frac{\partial Y}{\partial K}}{Y}$$

$$\eta_\ell = \frac{L \frac{\partial Y}{\partial L}}{Y}.$$

Under the assumption of constant returns to scale and a marginal productivity theory of value, η_k and η_ℓ can be obtained as empirical factor shares. Knowing growth rates of capital and labor, we can then calculate the factor growth contribution term $(\eta_k g_k + \eta_\ell g_\ell)$ for each year.

Now a legitimate question to ask is whether the factor growth contribution to equation (10) is "doing anything" more to explain the growth of output over and above what is explained by the obvious base case of a constant long-term trend modified by a short-term correction for economic activity. Is the sophisticated production function specification of potential output (7) statistically superior to the simple constant growth alternative

$$\hat{Y}(t) = B e^{\mu t} \tag{11}$$

which is frankly empirical?

In the language of this paper, <u>if</u> the true specification is

$$g_y = (\eta_k g_k + \eta_\ell g_\ell) + \lambda_1 + \gamma_1 g_e + \varepsilon \tag{12}$$

where ε is i.i.d. normal, how likely is it that a naive ad hoc specification which omits the production function part altogether

$$g_y = \lambda_2 + \gamma_2 g_e + \delta \tag{13}$$

should fit the data as well as it does? This way of posing the question seems like a fair way to inquire about the statistical likelihood of a stable production function relation between aggregate output, capital, and labor. Observe that (12) and (13) are <u>not</u> nested models in the usual statistical sense, although they are related to each other in a particularly

simple fashion through a zero-or-one coefficient.[2] Thus, our basic question about the role of the aggregate production function in explaining growth cannot be answered directly within the classical framework.

Note that we have imposed no conditions on the specific form of the production function F(K,L), having used only the assumption of competitive shares and constant returns to scale.

The data used are an updated and revised version of the Christianson - Jorgenson series on the aggregate private economy from the Review of Income and Wealth.[3] The primary period investigated was 1947-1978, although in all cases we have verified the legitimacy of our results over the longer span 1929-1978.

The ESS for the regression based on (12) is 7.072E-03(with a DW of 2.15) whereas the ESS is 7.024E-03 for the regression based on (13). Thus, omitting capital and labor altogether gives a slightly better fit than if the production function were included in the growth equation. More importantly, the difference is highly significant. The probability that (13) would fit as well as it does if (12) were the true specification is

$$\alpha = .06 \pm .01$$

(based on 1000 simulations). In this sense, we tend to reject the production function specification at a high level of significance.

Conclusion

Our guess is that the rejection of so-called "theoretically justified" economic models by empirical, ad-hoc specifications might be a pervasive feature in several (although by no means all) areas of applied econometric work.

Often what we deal with in economics is essentially a collection of time series, somewhat erratic but on the whole growing together. It is tempting to read into this mild chaos some order, and an economist naturally tries to impose order by fitting an economic model -- for example a production function. But we fear the sad truth is that many of our "theoretically justified" models may not be fitting the data much better than clever ad hoc specifications. And this would mean, in a sense we have tried to make rigorous, that the "theoretically justified" model is suspect on empirical grounds, however strong the desire or motivation to believe it is true.

Failing our test does not necessarily mean the hypothesis model must be abandoned. But it can signal us when to be more modest in our claims, more tentative in our applications, and more alert for alternative explanations.

We hope the exact test we have proposed can be constructively used to sift out those models where there exists an empirically viable structure from other situations where our own wishful thinking is making us impose a theoretical grouping that perhaps nature does not intend.

FOOTNOTES

[1]We checked these results empirically by performing several nested tests, and comparing the values of α obtained from a t or F distribution with our simulations:

Form of test	α from t or F dist.	α from simulations
$\beta_1 = 0$ (t)	0.640	0.625
$\beta_2 = 0$ (t)	0.776	0.760
$\beta_1=\beta_2=0$ (F)	0.884	0.890

The number of simulations in each case was 1,000.

[2]Of course models (12) and (13) could be artificially nested by introducing a free coefficient on the factor growth contribution term. However, this would not be the same test as we propose. In fact it is statistically much weaker because of the multicollinearity typically introduced by artificial nesting. The estimated coefficient is .34 with a standard error of .42.

[3]We thank Dale Jorgenson for providing us with this data.

REFERENCES

Cox, D.R.: "Tests of Separate Families of Hypotheses", Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, Berkeley: University of California Press, 1961.

Cox, D.R.: "Further Results on Tests of Separate Families of Hypotheses", Journal of the Royal Statistical Society, Series B, 24 (1962), 406-424.

Christensen, L.R. and Jorgenson, D.W. (1970) "U.S. Real Product and Real Factor Input, 1929-1967", Review of Income and Wealth, 16(1), March 19-50.

Davidson, R. and MacKinnon, J.G. (1980) "Several Tests for Model Specification in the Presence of Alternative Hypotheses", Discussion Paper no. 378, Queens University, forthcoming in Econometrica.

Pesaran, M.H.: "On the General Problems of Model Selection", Review of Economic Studies, 41 (1974), 153-171.

Pesaran, M.H. and A.S. Deaton: "Testing Non-Nested Nonlinear Regression Models", Econometrica, Vol. 46, No. 3 (May, 1978), 677-694.