TEACHING HYPOTHESIS TESTING: WHAT IS DOUBTED, WHAT IS TESTED?

Gordon Menzies
University of Technology, Sydney and Australian National University
E-mail:

ABSTRACT

Null hypotheses in undergraduate econometrics courses are usually framed in terms of parameter values or distributions. But relatively simple techniques can also test for violations of good scientific practice. This is neatly illustrated for students by a reinterpretation of an influential paper by Sir Ronald Fisher, where a rejection region is formed on the left tail of a χ^2 distribution. This idea is extended to situations where dubious models fit 'too well'. In these cases, a high R^2 may be taken as evidence that a non-random subset of regressions is being 'adversely selected' for publication.

Keywords: Teaching of econometrics; alternative hypothesis; Sir Ronald Fisher; R^2; Gregor Mendel; Adverse Selection.

***JEL Classification*:**

INTRODUCTION

In introductory econometrics courses, students learn about the logic of hypothesis tests. Rare events are those realizations of the test statistic which make us doubt a claim (the null Hypothesis) concerning a parameter value or distribution.

However, familiar statistical tests can also be used to test for data culling, or other improper scientific practices. This observation may be of interest to upper-year undergraduates.

1. THE LEFT-TAILED χ^2 TEST

We begin with a nice historical example, in which the independence of multinomial trials in a χ^2 goodness-of-fit test was successfully challenged by Sir Ronald Fisher. He could do so because the expected frequencies in the χ^2 test were based on Gregor Mendel's theory of genetics, which was widely accepted by Fisher's time. His insights are best presented in his own words:

> "Fictitious data can seldom survive a careful scrutiny, and, since most men [sic] underestimate the frequency of large deviations arising by chance, such data may be expected generally to agree more closely with expectation than genuine data would."
> (Fisher 1936, pp. 129-130)

In the same paper he showed that the calculated χ^2 goodness-of-fit test statistics for Gregor Mendel's plant hybridisation experiments were too small to be credible. This use of the χ^2 statistic led him to a startling conclusion.

> "… most, if not all, of the experiments have been falsified so as to agree closely with Mendel's expectations."
> (Fisher, 1936, p. 132)

Fisher implicitly ran a χ^2 goodness-of-fit test with the rejection region in the left hand tail.

[Fig. 1]

This is a test, where the null hypothesis is that the data is drawn from independent and identical multinomial trials (with the various success probabilities given by Mendel's theory). That is, the null is that the data is a true random sample. Fisher believed that Mendel's gardener had tampered with the data. While Fisher was not sure how the gardener did it, schemes that 'make the data fit' may be broadly described as a violation of independence,

since the tamperer must cull or alter observations bearing in mind the other (untampered) observations. He or she removes 'unfavourable' realizations, or even fabricates data. Either way, independence is violated.

In the χ^2-statistic, this violation of independence will clearly show up as a small value; hence the left-tail rejection region. The alternative hypothesis for this test of random sampling is that the distributions are not independent.

This test for random sampling is just as rigorous as a standard goodness-of-fit test. The goodness-of-fit test statistic has a χ^2 distribution (approximately) if the expected frequencies are correct, and the sampling is random (drawn from identical and independent multinomial trials). Both these conditions are necessary for the test statistic to have χ^2 distribution.

Therefore, depending upon what is doubted, the test statistic can be used as evidence for either non-random sampling or incorrect expected frequencies, but not both. If we are confident that random sampling has occurred, a high value of the test statistic convinces us (with probability α of making a mistake) that the expected frequencies are wrong. Similarly, if we are confident about the expected frequencies, a rejection region like the one in Figure 1 convinces us (with probability α of making a mistake) that random sampling has not occurred.[1]

This can be stated intuitively; the data may be used differently depending on what is doubted. If Mendel was not a party to the deception, what was in doubt was the frequency of plants with certain characteristics. When he saw the frequencies conforming (very closely) to his theory, his doubts were allayed. From Fisher's point of view, aided by subsequent scientific research which put the expected frequencies beyond reasonable doubt, the very same data provided conclusive evidence that the experiments were interfered with.

Fisher's approach circumvents (at the cost of making a type I error) the need for so called 'set-up experiments', where researchers suspected of fraud are surreptitiously placed in experimental environments that could not possibly produce the outcomes that the researchers claim they do. Such experiments have been criticized, because misconduct in the set-up experiment does not prove original misconduct (Office of Research Integrity, 2002).

The left-tailed χ^2 test could be used in other situations as well.[2] Consider the situation facing a project manager who has asked a researcher to collect a random sample from a population. Suppose further that the manager doubts that the sample was collected randomly, but instead suspects that it was collected to perfectly fit a demographic characteristic of the population, say an

[1] One can depart from random sampling in many ways; the particular departure in mind is massaging observed frequencies to agree with expected ones.

[2] I am grateful to staff at the Australian Bureau of Statistics with whom I have discussed this section.

age profile, but that it was otherwise poorly drawn. This suspicion could be confirmed by seeing if the data classified by age fitted 'too well' by using the left-tailed χ^2 test.

As was the case for Mendel's data, this use of the test requires a prior doubt in the integrity of the researcher. The exact same data could be used, with a rejection region in the right-hand-tail, if the concern was that the data drawn was not representative enough of the population. In that case the focus would not be on the researcher's integrity, but on other features of the sampling procedure.

2. INFLATED R^2 AND INVERTED F-TESTS

Researchers sometimes have the experience that they find an econometric model in a journal with a high R^2, despite having tried without success to model the relationship in question.

This could be explicable by adverse selection, a term coined in the insurance industry. The selection of people who purchase insurance is unlikely to be a random sample of the population. Instead, they are more likely to be a group with private information about their personal situations that makes them more likely to obtain a higher-than-average payout under the policy. Women contemplating pregnancy are more likely to take out a health policy with generous maternity provisions, confounding actuarial calculations based on the general population (Milgrom and Roberts 1992).

In the same way, the reported regressions arriving at the journal editor's desk might not be a random sample of regressions of that particular functional form in the population of regression realizations (i.e. for all sensible time periods and cross-sections). Instead, they might be sent by authors with private information about, say, all the modelling attempts that failed. Truly spurious models that look impressive (probably with a high R^2, among other things) might be overrepresented on the editor's desk.

Naturally, the claim that high-R^2 regressions might be 'adversely selected' out of the population of regressions is equivalent to claiming that regressions with a high F statistic are being adversely selected. As is well known, for the model $y_t=\beta_1+\beta_2x_{2t}+\ldots\beta_kx_{kt}+u_t$ the F-test for the null H_0: $\beta_2=\beta_3=..\beta_\kappa=0$ (i.e. $R^2=0$) can be written with R^2.

$$F=\frac{(RSS_{H0}-RSS_{H1})/J}{RSS_{H1}/(n-k)}=\frac{R^2/J}{(1-R^2)/(n-k)}\sim F_{J,n-k}.$$

Thus, a high R^2 is the same as a high F. However, *if one has solid grounds for believing that the purported relationship is not significant* (i.e. H_0: $\beta_2=\beta_3=..\beta_\kappa=0$), then a significant R^2 must be understood differently.

Now we describe the new hypotheses. The null would be that the regression in the journal is representative of the family of such attempted regressions (which use variables that are not, in fact, related). A significant R^2 leads to a rejection of that null (that the journal regression is representative), in favour of adverse selection.

H_0: A representative regression has been submitted to the journal

H_1: Journal 'adverse selection' has occured

In this case a significant R^2 leads to rejection of H_0.

That is to say, just as Mendel's gardener might have removed 'troublesome observations' so researchers might remove low-R^2 regressions. Naturally, this is not a useful perspective unless one has solid reasons to doubt the model, such as one's own fruitless attempts to model the relationship in question. Otherwise a high R^2 would always make one doubt the integrity of a researcher! The point is just that a significant R^2 can mean something quite different to what it is normally taken to mean, depending on what is doubted.

To drive home the similarity to the Mendel case, this test can even be cast as a left-tailed test (as Fisher's test was), though nothing hinges on this representation. Defining a new statistic as the inverse of the old one we have:

$$\mathrm{F}^{-1} = \frac{(1-R^2)/(n-k)}{R^2/J} \sim F_{n-k,J}.$$

And clearly as R^2 becomes arbitrarily close to one, the null of no adverse selection could be rejected on a left tail, for an arbitrarily small α.

[Fig. 2.]

The final example where a high R^2 raises suspicions concerns data averaging. We conceive of a situation that differs slightly from one outlined by the Office of Research Integrity (2002, pg. 16, Section 2), where an officer was asked to investigate a series of rat measurements. The data had one extra digit – a zero or five – compared with other measurements from the same experiment. The officer hypothesized that the series was the average of another two (zeros arising when the sum of the numbers had an even last digit, and fives when it was odd), and then confirmed that this was, in fact, the case.

But what would have happened if the tamperer had removed the last digit so that the series was not <u>exactly</u> the average of the other two? In this situation, the ghost of Fisher would urge us to run a multiple regression of each series suspected of being fabricated on the other series. If the regression included the two 'parent' series as independent variables, a very high R^2 (and coefficients very close to 0.5, or zero) would provide strong evidence of tampering. Naturally, this regression technique could also uncover fabricated

data using other kinds of linear combinations of existing data, such as averaging over more than two series.

3. PEDAGOGICAL CONSIDERATIONS

These ideas are suitable for presentation to students who have a very good grasp of the logic of hypothesis testing. Second-year econometrics students can think of examples of fabricated data, but they may not be comfortable enough with the standard meaning of a rare event to confront them with a non-standard meaning.[3]

Students really need to understand how hard it is to fabricate a random sample, before they can see the wisdom in the quote at the start of the paper by Sir Fisher. Less strong students could be led to being suspicious of, say, a perfect fit, and even weak ones could be shown how multiple regression could uncover averaging.

For the brightest students, or for upper-year undergraduates with a lot of experience of hypothesis testing, these examples can lead to a discussion of the importance of assumptions in hypothesis testing. It is impossible to test everything at once; Fisher regarded Mendels multinomial success probabilities as being accurate, and researchers suspicious of a high R^2 must have reasons to doubt the purported relationship. Out of the range of possible doubts, what is doubted *specifically* determines what is tested.

[3] I ran an experiment in an introductory regression class to see if respondents could tell if a cup of tea had had the milk, or tea, poured in first (following a famous experiment at Cambridge University). Out of 49 students observing the experiment, only 7 were suspicious when I asked them to comment on (fabricated) data from an alleged similar experiment where someone got 100 correct identifications from 100 cups of tea.

REFERENCES

Fisher, R. A. (1936), "Has Mendel's Work Been Rediscovered," *Annals of Science*, Vol. 1 (2) April, 115-137.

Milgrom P. and J. Roberts (1992), *Economics, Organization and Management*, Prentice Hall, Upper Saddle River, New Jersey.

Office of Research Integrity (2002), "Annual Report 2001" Dept. of Health and Human Service Office to the Secretary Office of Public Health and Science, July 2002.

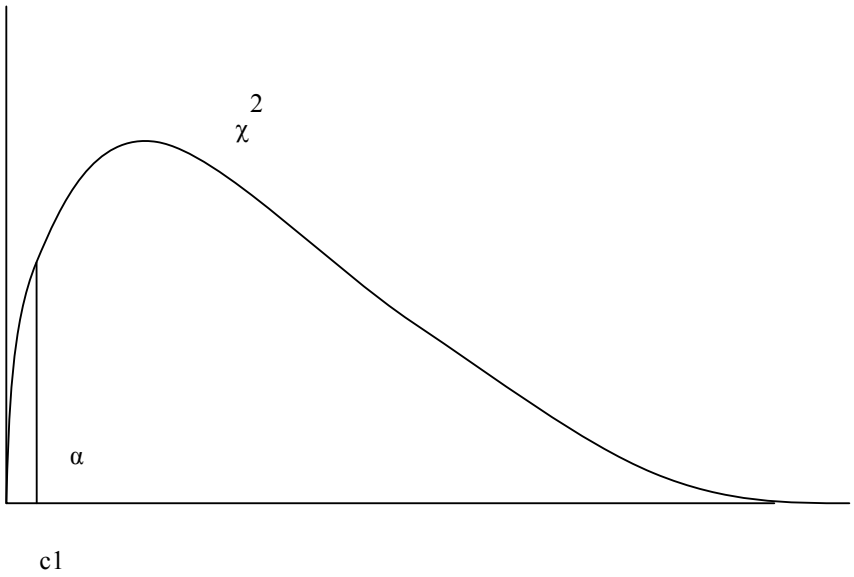

Figure 1. Fisher's logic motivates a hypothesis test. Fisher showed that Mendel's observed frequencies were too close to the theoretically expected frequencies by calculating a χ^2-goodness-of-fit test statistic. The null hypothesis is random sampling, while the alternative hypothesis is that the data has been tampered with to make the observed frequencies close to the expected frequencies. A value less than c_1 is evidence that independence among the data has been violated. If the sampling is random, the probability of this occurring by chance is α.

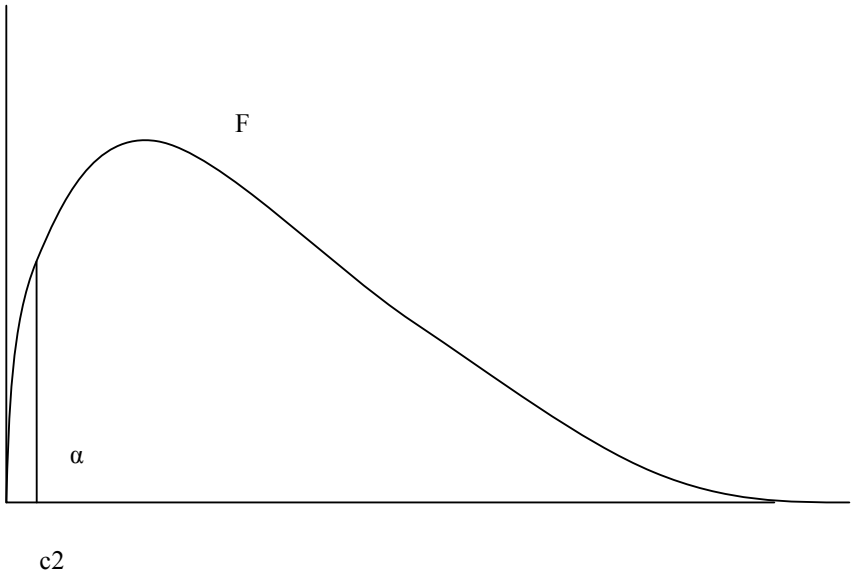

Figure 2. A left-hand-rejection-region is appropriate for an (inverted) F-test too. The null hypothesis is no adverse selection. Yet for very high values of the reported R^2, such that the inverted F statistic falls below c_2, you believe instead that a very unrepresentative regression has made its way to the journal editor's office. If you are wrong about this, the probability of making a mistake is α.