econometrics

Article

The Replication Crisis as Market Failure

John Quiggin

School of Economics, The University of Queensland, St Lucia QLD 4072, Australia; j.quiggin@uq.edu.au

Received: 30 September 2019; Accepted: 18 November 2019; Published: 24 November 2019

Abstract: This paper begins with the observation that the constrained maximisation central to model estimation and hypothesis testing may be interpreted as a kind of profit maximisation. The output of estimation is a model that maximises some measure of model fit, subject to costs that may be interpreted as the shadow price of constraints imposed on the model. The replication crisis may be regarded as a market failure in which the price of "significant" results is lower than would be socially optimal.

Keywords: replication crisis; profit maximization; market failure

1. Introduction

Concerns about the inadequacy of standard practices in statistics and econometrics have been long-standing. Since the 1980s, criticisms of econometric practice, including those of Leamer (1983); Lovell (1983); McCloskey (1985) have given rise to a large literature. Kim and Ji (2015) provide a survey. Even more significant, within the economics profession was the Lucas (1976) critique of Keynesian economic modelling Lucas and Sargent (1981).

Most of the concerns raised in these critiques apply with equal force to other social science disciplines and to fields such as public health and medical science. Ioannidis (2005) offered a particularly trenchant critique, concluding that "most published research findings are false".

The emergence of the "replication crisis", first in psychology and then in other disciplines, has attracted the broader public to some of these concerns. The applicability of the term "crisis" is largely due to the fact that, unlike previous debates of this kind, concern over replication failures has spilled over disciplinary boundaries and into the awareness of the educated public at large.

The simplest form of the replication crisis arises from the publication of a study suggesting the existence of a causal relationship between an outcome of interest y and a previously unconsidered explanatory variable x followed by studies with a similar design that fail to find such a relationship.

Most commonly, this process takes place in the context of classical inference. In this framework, the crucial step is the rejection of a null hypothesis of no effect, with some specified level of significance, typically 95 per cent or 90 per cent. In the most commonly used definition, a replication failure arises when a subsequent study testing the same hypothesis with similar methods but with a different population fails to reject the null.[1]

For example, Kosfeld et al. (2005) found that exposure to oxytocin increases trust in humans. This finding created substantial interest in possible uses of oxytocin to change mood, potentially for malign as well as benign purposes. Similar results were published by Mikolajczak et al. (2010). However, a subsequent attempt at replication by the same research team, Lane et al. (2015), was unsuccessful.

[1] As an anonymous referee points out, this characterisation of replication failure is too stringent since some failures to reject the null are to be expected. More stringent definitions of replication failure are that the null is not rejected using the pooled data or that the parameter estimates from the two studies are (statistically) significantly different.

The crisis was brought to the wider public's attention by the publication by Open Science Collaboration (2015) of a systematic attempt to replicate 100 experimental results published in in three psychology journals. Replication effects were half the magnitude of original effects, representing a substantial decline. Whereas ninety-seven percent of original studies had statistically significant results, only thirty-six percent of replications did.

A variety of responses have been offered in response to the replication crisis. These include tightening the default P-value threshold to 0.005 (Benjamin et al. 2018), procedural improvements such as the maintenance of data sets and the preregistration of hypotheses ((Nosek and Lakens 2014), attempts to improve statistical practice within the classical framework, for example, through bootstrapping (Lubke and Campbell 2016)), and the suggestion that Bayesian approaches might be less vulnerable to these problems (Gelman 2015).

This paper begins with the observation that the constrained maximisation central to model estimation and hypothesis testing may be interpreted as a kind of profit maximisation. The output of estimation is a model that maximises some measure of model fit, subject to costs that may be interpreted as the shadow price of constraints imposed on the model. This approach recalls the observation of Johnstone (1988) "That research workers in applied fields continue to use significance tests routinely may be explained by forces of supply and demand in the market for statistical evidence, where the commodity traded is not so much evidence, but "statistical significance."[2]

In mainstream economics, an unsatisfactory market outcome is taken as prima facie evidence of a "market failure", in which prices are not equal to social opportunity costs.[3]

In this paper, we will consider the extent to which the replication crisis, along with broader problems in statistical and econometric practice, may be seen as a failure in the market that generates published research.

2. Model Selection as an Optimisation Problem

In the general case, we consider data (X, y) where y is the variable (or vector of variables) of interest, and X is a set of potential explanatory variables. We consider a finite set of models $\mathcal{M}$, with typical element m. The set of models may be partitioned into classes $\mathcal{M}_\kappa$, where $\kappa = 1 \ldots K$. Typically, although not invariably, lower values of κ correspond to more parsimonious and, therefore, more preferred models.

For a given model m, the object of estimation is to choose parameters $\beta^* (m; X, y)$ to maximise a value function $V(\beta; X, y)$, such as log likelihood or explained sum of squares. Define

$$V^* (m; X, y) = \max_\beta V(\beta; X, y) = V(\beta^* (m; X, y); X, y). \quad (1)$$

The model selection problem is to choose m to maximise the global objective function

$$\Pi(m; X, y) = V^* (m; X, y) - C(m), \quad (2)$$

where $C(m)$ is a cost function. Given the interpretation of V as a value function and C as a cost function, Π may be regarded as a profit function.

[2] I am indebted to an anonymous referee for pointing out this article, foreshadowing my central point.

[3] Quiggin (2019) extends this analysis to encompass issues such as unemployment and inequality.

2.1. Linear Case

We will confine our attention to linear models with a single variable of interest y and N observations on a set of K potential explanatory variables $X = (x_1...x_K)$. The generic model of this form is

$$\begin{aligned} Y &= X\beta + \varepsilon \\ R\beta &= v, \end{aligned} \tag{3}$$

where

Y is an $N \times 1$ vector of observations on y;
$\mathbf{X}$ is an $N \times K$ matrix of observations on X;
β is a $K \times 1$ vector of parameters;
ε is an $N \times 1$ error term;
R is a $J \times K$ vector of constraints, where $J < K$ and R has full rank;
the special case of ordinary least squares (OLS) is that of k unconstrained explanatory variables. In this case, $J = K - k$ and $R = \begin{pmatrix} \mathbf{0}_k \\ \mathbf{I}_{K-k} \end{pmatrix}$. The model may be written without explicit constraints as

$$Y = \mathbf{X}\beta + \varepsilon, \tag{4}$$

where

Y is an $N \times 1$ vector; of observations on y
$\mathbf{X}$ is an $N \times k$ matrix of observations on $(x_1...x_k)$;
β is a $k \times 1$ vector of parameters;
ε is an $N \times 1$ error term, distributed as iid.$N(0, \sigma)$

2.2. Value Functions, Cost Functions, and Profit Functions

2.2.1. Value

The most commonly used value measures are
(i) measures related to likelihood,

$$\mathcal{L} = \prod_n p\left(y_n | \beta\right); \tag{5}$$

(ii) those based on the proportion of variation in y explained by the model. The canonical measure is

$$R^2 = \frac{RSS}{TSS}. \tag{6}$$

These value measures may be interpreted in terms of explanation and prediction. If the objective of the model is taken to be the explanation of observed outcomes in terms of a given model, the best explanation may be seen as the choice of β that maximises the likelihood of the observed y. If the objective of the model is taken to be prediction of y given X, R^2 is the value function implied by the minimisation of a quadratic loss function.

2.2.2. Cost

The simplest cost functions C are monotonic functions of k, the number of non-zero parameters in the model. More generally, we may attach a cost c to the relaxation of a constraint $R\beta = 0$ suggested by theory.

Any cost function gives rise to a partition of the set of models into classes $\mathcal{M}_\kappa$ where $\kappa = 1 \dots K$, where the equivalence relation is defined by the property that if $\kappa(m) = \kappa(m')$, $C(m) = C(m')$.

Standard choices include

$$\begin{aligned} C(m) &= k, \\ C(m) &= \frac{k-1}{N-k-1}, \\ C(m) &= K - J. \end{aligned} \tag{7}$$

2.2.3. Profit

In a general setting, profit may be defined as

$$\Pi = py - wx, \tag{8}$$

where y is an output measure, p is the output price, x is an input measure, and w is the wage or factor price. Treating output as numeraire, we may simplify to

$$\Pi = y - wx. \tag{9}$$

Given the interpretation above, profit maximisation is represented by the choice of the best-fitting model, subject to a cost associated with rejecting restrictions suggested by theory or by concern about Type 1 error.

As in a market setting, profit maximisation will be consistent with a broader notion of optimality if y and x are measured correctly and if the price w represents social cost.

3. Model Selection and Hypothesis Testing

To examine problems of hypothesis testing and model selection, we introduce an adjacency relationship between models. Informally, two models are adjacent if they differ by a single step such as the inclusion of an additional variable or the imposition of a linear restriction. As these examples indicate, it will typically be the case that adjacent models m, m' are ranked by parsimony so that, if m is derived by imposing a restriction in m', then m' $\kappa(m) < \kappa(m')$, and there exists no m'' such that $\kappa(m) < \kappa(m'') < \kappa(m')$. The adjacency relation is directed and will be denoted by $m \rightarrow m'$ (for the case $\kappa(m) < \kappa(m')$).

3.1. Global Model Selection Criteria

Many widely used model selection criteria may be interpreted as profit functions, with a reversal of sign to make the problem one of maximisation rather than minimisation

Examples include the Akaike information criterion (AIC),

$$-AIC = 2\log(\mathcal{L}) - 2k, \tag{10}$$

the corrected AIC,

$$-AICc = AIC - \frac{2k^2 + 2k}{N-k-1}, \tag{11}$$

and the Bayesian information criterion (BIC)

$$-BIC = 2\log(\mathcal{L}) - \log(N)\,k. \tag{12}$$

It is also possible to include a cost function in measures based on R^2.

For example, we may write

$$V = R^2 - \frac{k-1}{N-k-1}. \tag{13}$$

Closely related is the $\bar{R}^2$ criterion, which satisfies

$$1 - \bar{R}^2 = (1 - R^2) \frac{N-1}{N-k-1} \tag{14}$$

or

$$\bar{R}^2 = \frac{ESS/\left(k-1\right)}{RSS/\left(N-k-1\right)}. \tag{15}$$

3.2. Local Model Selection Criteria

Definition 1. *Let $m \to m'$ be adjacent models. Then m' is a stepwise profit improvement (reduction) over m if $\Pi\left(m'; X, y\right) > (<) \, \Pi\left(m'; X, y\right)$.*

This definition immediately suggests consideration of the stepwise regression algorithm proposed by Efroymson (1960). Let the profit function Π be $\bar{R}^2$. From an initial m, choose the adjacent m' that maximises $\Pi\left(m'; X, y\right)$ terminating when, for all adjacent $m'', \Pi\left(m'; X, y\right) \geq \Pi\left(m''; X, y\right)$. Forward selection adds the requirement $m \to m'$ so that selection proceeds by adding variables. Conversely, backward selection requires $m' \to m$.

3.3. Hypothesis Testing

The problems of model selection and hypothesis testing are commonly treated separately. From the viewpoint offered here, they may be regarded as global and local versions of the same problem, that of profit maximisation. This point may be illustrated with respect to the large class of classical hypothesis tests encompassed by the Wald, likelihood ratio, and Lagrange multiplier statistics Engle (1984), and also with respect to the t and F tests used in hypothesis testing in the OLS framework.

Model selection may be considered locally in terms of a pairwise choice between models, so that m is preferred to m' if and only if $\Pi\left(m; X, y\right) > \Pi\left(m'; X, y\right)$, which may be written as

$$V^*\left(m; X, y\right) - C\left(m\right) > V^*\left(m'; X, y\right) - C\left(m'\right) \tag{16}$$

or

$$V^*\left(m; X, y\right) - V^*\left(m'; X, y\right) > C\left(m\right) - C\left(m'\right). \tag{17}$$

If $V\left(m; X, y\right) = \log\left(\mathcal{L}\right)$, the left-hand side of (16) and (17) is a log likelihood ratio. If m' is obtained from m by the imposition of a vector of J constraints $R\beta = v$, then under the standard assumptions of the general linear model, the Likelihood Ratio (LR) statistic is distributed as χ^2 with J degrees of freedom in the limit. Hence, if we set $C\left(m'\right) = 0$ and $C\left(m'\right)$ equal to a suitably chosen critical value for $\chi^2\left(J\right)$, we obtain the usual LR test.

Compare this approach to the various information criteria discussed in the previous section. The domains of the two approaches coincide in the case of nested models where the restrictions imposed in m' consist of setting some coefficients equal to zero. The canonical criteria set out above have in common that more parsimonious models are always preferred, at least weakly. In formal terms, the classical hypothesis-testing framework shares this characteristic. The only difference is in the choice of cost function.

The same point may be made with respect to value functions based on explained variation. The natural starting point is the F-test,

$$F = \frac{\left(RSS\left(m\right) - RSS\left(m'\right)\right)/J}{RSS\left(m\right)/\left(N-k\right)}. \tag{18}$$

Alternatively, we may consider the framework of constrained optimisation. The objective for a linear model m may be written as

$$V\left(m; X, y\right) - \lambda\left(R\beta - v\right), \tag{19}$$

where λ is a vector of Lagrange multipliers, such that $\lambda\left(R\beta - v\right) = 0$. In the context of constrained optimisation, as observed by Breusch and Pagan (1980), the Lagrange multiplier represents the shadow price of relaxing a constraint.

The most usual restriction is that of setting the coefficient on some variable equal to zero. In this case, the F statistic is typically replaced by its square root, and the test is a t-statistic.

Such tests of significance give rise to an estimation strategy in which variables are included sequentially in the model. The most usual criterion is the t-statistic associated with the added variable, typically with a stopping criterion associated with 5 per cent significance.

As Dhrymes (1970) observes, this strategy is closely related to the criterion of maximising $\bar{R}^2$. However, $\bar{R}^2$ is (locally) optimised by including variables whenever their t-statistic is greater than one and, therefore, implies a lower price for including variables. A similar point may be made with respect to the relationship between information criteria such as the AIC and the likelihood ratio test.

4. Market Failure

When the model selection problem is interpreted as one of profit maximisation, it is natural to interpret problems within econometric practice as market failures. In particular, the profit function being maximised by researchers differs from that which would be suggested by an objective of social welfare maximisation.

In relation to the replication crisis, the core of the problem, as it is generally perceived, is that the price of including a variable of interest as statistically or economically significant in the reported model is too low.

On the one hand, the bias against publication of negative results means that the benefit to researchers of reporting models with no significant variables of interest is limited and, in many fields, close to zero.

On the other hand, the availability of the set of techniques pejoratively referred to as "p-hacking" means that the test statistics reported in published studies are more likely to arise through chance than would be suggested by the classical hypothesis-testing framework.

Taken together, these observations suggest that the price to researchers of including a variable of interest in a published model is lower than would be socially optimal. As a result, too many positive results are reported.

But what is the socially optimal price? The classical hypothesis-testing framework is of limited value here. Informal reasoning about Type 1 and Type 2 errors implies that there must exist a trade-off that can be expressed in terms of relative prices. But the relationship between that trade-off and the notion of statistical significance is opaque, to say the least. Holding the size (likelihood of Type 1 error) constant means that the power (likelihood of Type 2 error) depends on a combination of sample size and the variance of the error term in a way that bears no obvious relationship to the relative desirability of avoiding the two types of error.

Bayesian approaches based on loss functions are much closer to the spirit of the approach being suggested here. The main difference between the profit maximisation approach proposed here and the loss function approach (apart from a trivial change of sign) is the inclusion of an explicit cost associated with the estimation of a more general model.

The central point of the market failure analogy is that excessive publication of fragile results implies that the price of a positive result is too low. As with other market failures, a solution may be sought either in regulation or in pricing.

Regulatory approaches include measures such as the preregistration of estimation strategies. To develop an appropriate model of socially optimal prices, we must consider how, as a society, we respond to research publications.

This is, in some sense, an equilibrium outcome. If we accept at face value either the classical interpretation of a finding of statistical significance or its natural subjective misinterpretation (the probability that the variable in question is causally related to the variable of interest is near one), then the fact that many findings cannot be replicated is a major social problem. The initial result would lead us to act on a belief that cannot be substantiated.

Such an outcome cannot be sustained, however. Most people now understand that a reported result, whatever the supposed statistical significance, is not conclusive in the absence of replication, and certainly not at the stated level of significance. The question, therefore, is whether the publication of the result is, on balance, socially beneficial.

In this view, the most important requirement is that the price should be high enough to offset the inevitable effects of publication bias. Publication of a result showing a statistically significant result should be treated as an indication that further research is warranted, rather than conclusive or even highly probable evidence that the reported relationship is real.

5. Concluding Comments

Statistical research is a social and economic enterprise aimed at discovering, testing, and ultimately acting on relationships between variables of interest. The economic concepts of profit maximisation and market failure provide a way of thinking about statistical research that reflects its social role.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflicts of interest.

References

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E. -J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, and Colin Camerer. 2018. Redefine statistical significance. *Nature Human Behaviour* 2: 6–10. [CrossRef] [PubMed]

Breusch, Trevor S., and Adrian R. Pagan. 1980. The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics. *Review of Economic Studies* 47: 239–53. [CrossRef]

Dhrymes, Phoebus. J. 1970. On the Game of Maximizing R Bar Square. *Australian Economic Papers* 9: 177–85. [CrossRef]

Efroymson, M. 1960. Multiple regression analysis. In *Mathematical Methods for Digital Computers*. Edited by A. Ralston and H. S. Wilf. New York: Wiley.

Engle, Robert F. 1984. Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of Econometrics* 2: 775–826.

Gelman, Andrew. 2015. The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective. *Journal of Management* 41: 632–43. [CrossRef]

Ioannidis, John PA. 2005. Why most published research findings are false (Essay). *PLoS Medicine* 2: e124. [CrossRef] [PubMed]

Johnstone, David. 1988. Comments on Oakes on the foundation of statistical inference in the social and behavioral sciences: The market for statistical significance. *Psychological Reports* 63: 319–31. [CrossRef]

Kim, Jae H., and Philip Inyeob Ji. 2015. Significance testing in empirical finance: A critical review and assessment. *Journal of Empirical Finance* 34: 1–14. [CrossRef]

Kosfeld, Michael, Markus Heinrichs, Paul J. Zak, Urs Fischbacher, and Ernst Fehr. 2005. Oxytocin increases trust in humans. *Nature* 435: 673. [CrossRef]

Lane, Anthony, Moïra Mikolajczak, Evelyne Treinen, Dana Samson, Olivier Corneille, Philippe de Timary, and Olivier Luminet. 2015. Failed Replication of Oxytocin Effects on Trust: The Envelope Task Case. *PLoS ONE* 10: e0137000. [CrossRef]

Leamer, Edward E. 1983. Let's Take the Con Out of Econometrics. *The American Economic Review* 73: 31–43.

Lovell, Michael 1983. Data mining. *Review of Economics and Statistics* 45: 1–12.

Lubke, Gitta H., and Ian Campbell. 2016. Inference Based on the Best-Fitting Model Can Contribute to the Replication Crisis: Assessing Model Selection Uncertainty Using a Bootstrap Approach. *Structural Equation Modeling: A Multidisciplinary Journal* 23: 479–90. [CrossRef] [PubMed]

Lucas, Robert E., Jr. 1976. Econometric Policy Evaluation: A Critique. In *Carnegie-Rochester Conference Series on Public Policy*. Edited by Karl Brunner and Alan Meltzer. Amsterdam: North-Holland.

Lucas, Robert E., and Thomas J. Sargent. 1981. *Rational Expectations and Econometric Practice*. London: George Allen & Unwin.

McCloskey, Donald N. 1985. The loss function has been mislaid: The rhetoric of significance tests. *American Economic Review* 75: 201–5.

Mikolajczak, Moïra, James J. Gross, Anthony Lane, Olivier Corneille, Philippe de Timary, and Olivier Luminet. 2010. Oxytocin Makes People Trusting, Not Gullible. *Psychological Science* 21: 1072–74. [CrossRef] [PubMed]

Nosek, Brian A., and Daniël Lakens. 2014. Registered reports: A method to increase the credibility of published results. *Social Psychology* 45: 137–41. [CrossRef]

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349: aac4716. [CrossRef] [PubMed]

Quiggin, John. 2019. *Economics in Two Lessons: Why Markets Work So Well, and Why They Can Fail So Badly*. Princeton: Princeton University Press.