Word Count (including references and notes): 5480

On Goodness of Fit

Jonathan Livengood
University of Illinois at Urbana-Champaign

Abstract. We want our theories to be constrained by the data. We want them to have good fit. Statistical procedures that rely on a measure of how well a model fits some data are ubiquitous in the sciences. But there are many possible measures of fit. Why choose one measure of fit over its competitors? One might hope to find a purely epistemic justification for one's choice of fitting function. But I will argue that no purely epistemic reasons for choosing a fitting function can be given. We choose to adopt the fitting functions we do, not because they are epistemically justified, but because they are pragmatically attractive.

Keywords: goodness of fit, model, accuracy, pragmatic encroachment

Acknowledgements: [[Kindnesses go here.]]

On Goodness of Fit

Not so very long ago, I was talking about statistics with a friend from the psychology department. “Everything is regression,” my friend told me. Here is what I think she had in mind. Whenever we collect data in order to test a theoretical model, we measure the *fit* of the model to the data. When we assess the predictive accuracy of a model, we measure how well the model’s predictions fit our observations. And when we select a specific model from a family, we select the one that best fits the observations. We want our theories to be constrained by the data. We want them to have good fit. Statistical procedures that rely on a measure of how well a model fits some data are ubiquitous in the sciences. But there are many possible measures of fit. Why choose one measure of fit over its competitors? One might hope to find a purely epistemic justification for one’s choice of fitting function. But I will argue that no purely epistemic reasons for choosing a fitting function can be given. We choose to adopt the fitting functions we do, not because they are epistemically justified, but because they are pragmatically attractive.

Consequently, statistical procedures used across the sciences—from the hardest of the physical sciences to the softest of the human sciences—are pragmatically encroached. And as a result, science itself stands on practical feet of clay. So as to be clear up front, my basic argument goes like this:

[L1] Every plausible reason for choosing a measure of fit is pragmatic.

[L2] If [L1], then statistical procedures used across the sciences are pragmatically encroached.

[L3] If statistical procedures used across the sciences are pragmatically encroached, then science is pragmatically encroached.

[L4] Science is pragmatically encroached.

Here is how I proceed. In Section 1, I introduce two simple, historically interesting measures of fit. In Section 2, I motivate [L1] by considering some reasons commonly offered by statisticians and related researchers for choosing a measure of fit. In Section 3, I offer a principled argument for [L1] based on the idea that any epistemic argument for choosing a measure of fit would be viciously circular. In Section 4, I argue for [L2]. In brief, if [L1] is correct, then every instance of estimation, testing, or model selection carries with it some pragmatic commitment. Hence, for any statistical procedure, there will be some change in pragmatic commitments that will change its deliverances. In Section 5, I argue for [L3]. Science rides on the back of estimation, testing, and model selection. Hence, science cannot get very far without pragmatic commitments. Finally, in Section 6, I make some concluding remarks and compare my position with Douglas' (2000) account of inductive risk and values in science.

1. Groundwork

Two simple approaches to measuring goodness of fit have figured prominently in the history of statistics: least absolute error and least square error.[1] According to the least absolute error approach, the fit of a model is assessed by summing the absolute values of the differences between the observed data-values and the values delivered by the model. A model has better fit than a competing model if the sum of its absolute errors is smaller than the sum of the absolute errors of its competitor. According to the least square error approach, the fit of a model is assessed by summing the squared values of the differences between the observed data-values and the values delivered by the model. A model has better fit than a competing model if the sum of its squared errors is smaller than the sum of the squared errors of its competitor.

[1] See Portnoy & Koenker (1997) for an introduction to the statistical debate. For historical details in the linear case, see Farebrother (1999). For a nice comparison of some more-recent alternatives as applied to generalized curve-fitting problems, see Chapter 9 of Birkes & Dodge (2011).

Now, consider a simple curve-fitting problem. Suppose we have a two-dimensional scatterplot of data points, and we want to fit a straight line to the data.[2] The best-fitting model according to the standard of least absolute error may differ in important ways from the best-fitting model according to the standard of least square error: especially when there are outliers or when the noise has a distribution with heavy tails. Figure 1 shows a concrete example using data on income inequality and infant mortality.[3]

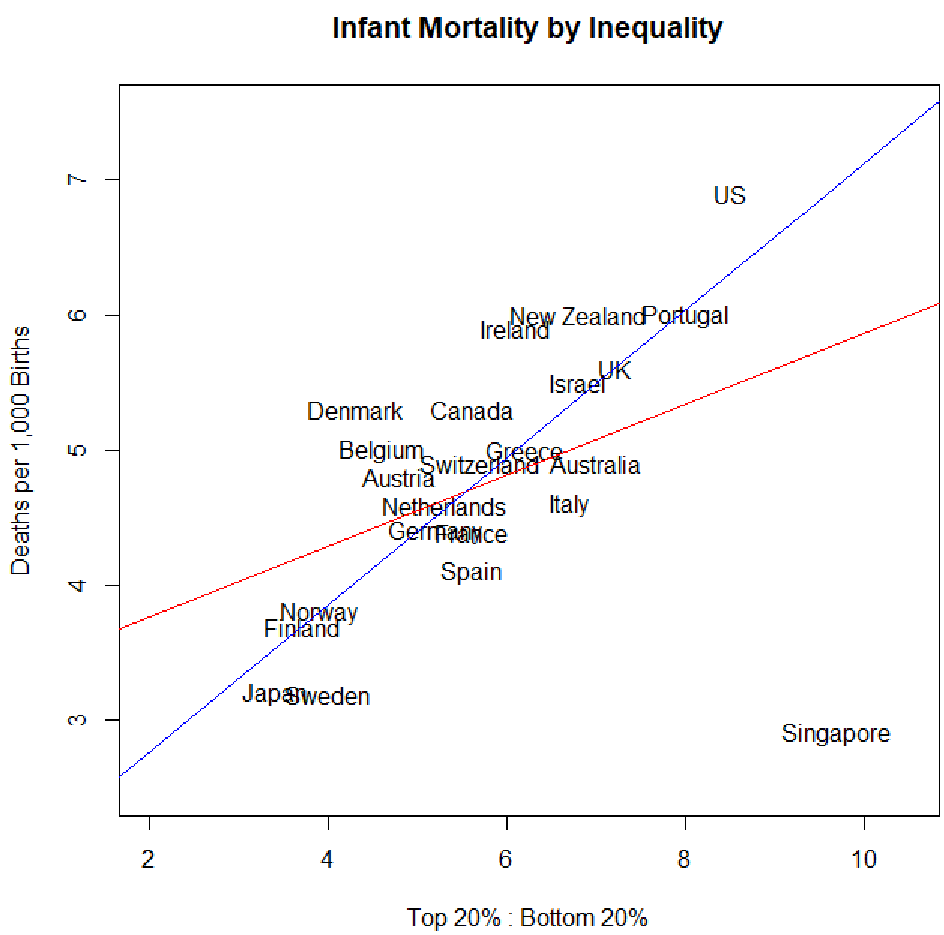

Figure 1

[2] See Glymour (1980), Forster & Sober (1994), and Spanos (2007) for discussions of the problem of selecting the appropriate family of curves in problems like this one.
[3] For a more politically neutral example, see Prescott (1975).

The blue line represents the preferred linear model as judged by the standard of least absolute error, and the red line represents the preferred linear model as judged by the standard of least square error. Singapore, with its extreme inequality and universal healthcare is an outlier that exerts considerable influence on the least squares error line (in red) but very little influence on the least absolute error line (in blue).

In estimation tasks for one-dimensional quantities, like estimating the speed of light or the charge of an electron, the least absolute error approach is equivalent to taking the median and the least square error approach is equivalent to taking the mean. As with the curve-fitting example, the absolute error approach (median) is less sensitive to outliers than is the square error approach (mean). If our estimates only have to be very approximately correct, then the difference between using the mean and using the median might not matter. But if we need to be very precise, the differences will often be meaningful.

Here is an illustration. In 1882, Simon Newcomb used the method of Foucault to make several measurements of the time that light took to travel a distance of 7,442 meters.[4] In Figure 2, I have reproduced Newcomb's measurements, which are given by replacing M in the expression $M \times 10^{-3} + 24.8$ with the values in my figure, where the result is in units of millionths of a second.

[4] See Stigler (1977) for Newcomb's data (in Table 5 of the appendix) along with a fascinating discussion of applying different estimation techniques to old data sets.

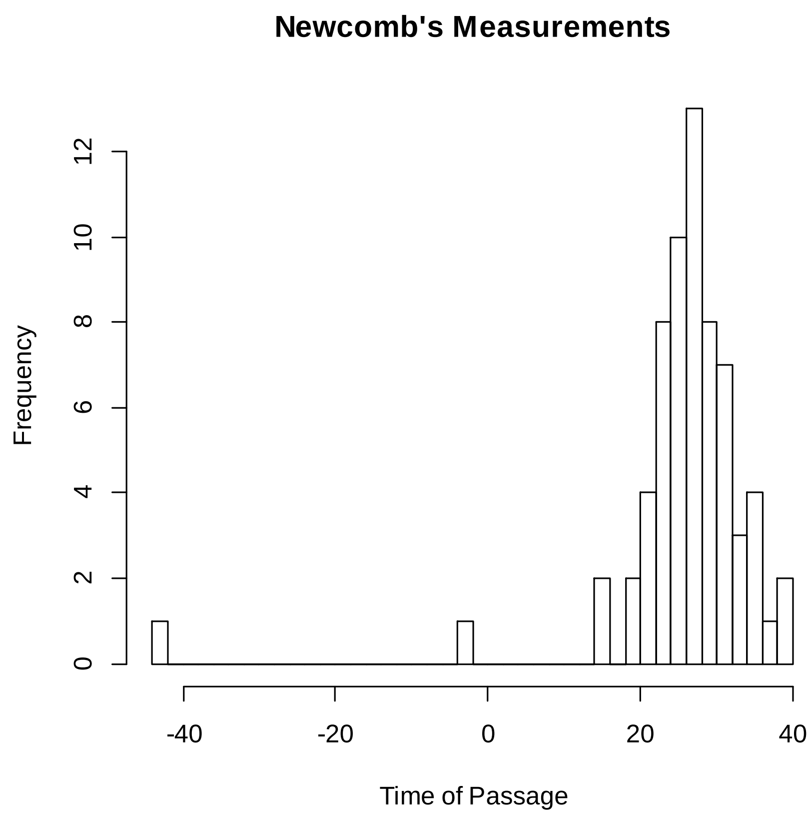

Figure 2

Newcomb used his measurements of the time of passage to estimate the speed of light. Using the mean (26.2), we get an estimate for the speed of light equal to 299,763,813 meters per second. Using the median (27), we get an estimate of 299,754,300 meters per second. The difference is surely trivial for most applications, but for some applications, such as estimating the output of a nuclear reaction or carrying out radio communications with deep space probes doing delicate, time-sensitive operations, the difference might be important.

Model selection techniques offer another example. An information criterion, such as the Akaike information criterion (AIC) or the Bayes information criterion (BIC), can be expressed as the sum of a term giving the measure of fit and a term penalizing a model based on its degree of complexity. Both the AIC and the BIC use square error in the goodness of fit term. They differ in the penalty term, with AIC using the penalty term $2k/_n$ and BIC using the term $k \cdot ln(n)/_n$, where *k* is the number of free parameters in the model and *n* is the sample size. If we use the same penalty terms with an absolute error fitting function, we get different results in some cases. Consider an application to Hald's cement data (a standard test-case for model selection techniques). The problem is to model the heat evolved in the hardening of cement as a linear function of the percentage weight of various chemicals that the cement contains. The Hald data set includes four predictor variables, denoted here by X_1, X_2, X_3, and X_4, where:

X_1 is the percentage weight of $3CaO.Al_2SO_3$.
X_2 is the percentage weight of $3CaO.SiO_2$.
X_3 is the percentage weight of $4CaO.Al_2O_3.Fe_2O_3$.
X_4 is the percentage weight of $2CaO.SiO_2$.

If one uses least square error with the AIC penalty term in the usual way, the model selected is $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \varepsilon$. And if one uses least square error with the BIC penalty term in the usual way, the model selected is $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. But if one uses least *absolute* error, the model selected is $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, regardless of whether one uses the AIC penalty term or the BIC penalty term.

2. Reasons for Selecting a Measure of Model Fit

What reasons might one give for preferring one measure of model fit over another? In this section, I consider four reasons given in the statistics and econometrics literatures: (1) the fitting function has some mathematical or computational virtues that its competitors lack; (2) the fitting function is the unique function that minimizes one's losses in the long run; (3) the fitting function is the most *efficient* one; and (4) the fitting function is more *robust* than its competitors to model misspecification. In this section, I argue that each of these reasons is pragmatic, not epistemic.

2.1 Mathematical or Computational Tractability

One appealing constraint on the choice of fitting function is mathematical or computational tractability. Historically, least square error was preferred because the optimization problem one faces when one uses least square error has an analytic solution; whereas, the problem one faces when one uses least absolute error requires numerical methods. More recently, the discussion has shifted to computational tractability, since computers solve both sorts of problems using numerical methods.

But what connection is there between our practical ability to solve a problem and the resulting solution's closeness to the truth? Does our having an easier time generating a solution confer greater *epistemic justification* on the solution? Prima facie, the answer is no. The fact that we have an easier time computing some function does not make it epistemically superior. One might try to resist this initial impression by suggesting that calculations having significantly less difficulty are correspondingly more epistemically secure. We are more likely to get the correct answer if the calculation is easy than if it is difficult. However, the reply is misguided. Correctly computing a function is not the same thing as computing a correct function.

Alternatively, one might argue that by using a computationally efficient method, one may get close to the truth for many different problems; whereas, using an inefficient method would mean that one misses out on the truth about *other* problems. I see two difficulties with this line of reply. First, there does not seem to be any way to say how close to the truth is close enough without knowing the purpose of solving the problem. But the point of solving a problem—the goal being set—is a pragmatic consideration. Hence, adopting an efficient method in order to get close to the truth on several problems presupposes a pragmatic criterion. The idea of getting close enough to the truth is by its nature to get close enough for practical purposes. Second, since closeness to the truth is exactly the notion under consideration, one cannot appeal to it without vicious circularity. Consequently, my imagined critic would need to replace the goal of getting close to the truth on many different problems with something like, “Yielding more answers that have good-making feature [x].” But what could play the role of good-making feature [x] is not obvious.

2.2 Loss Minimization

When one faces a problem involving estimation, an appealing constraint on the choice of fitting function (and by extension, estimator) is that one’s choice should minimize expected losses in the long run. Different loss functions (which are, effectively, utility functions) correspond to different fitting functions. Least square error is preferable if one has a squared error loss function; whereas, least absolute error is preferable if one has an absolute error loss function. But loss functions have to do with an agent’s aims and attitudes towards different varieties of risk or error. Considerations of risk are quintessentially pragmatic considerations. Hence, choosing a fitting function on the basis of one’s loss function is to choose a fitting function for a pragmatic reason.

Two agents who come to a problem with different loss functions (and hence, solve the goodness of fit problem in different ways) face a disagreement that cannot be settled by appeals to shared epistemic norms or by appeals to additional data. Someone who endorses the square error fitting function cannot be persuaded on the basis of data to move to absolute error. By her own lights, her model—fit according to her measure of goodness of fit—is as good as is possible. By her opponent’s lights, some *other* model has better fit. If their disagreement has any content at all, it is in virtue of some difference in their aims or desires. Hence, the source of their disagreement must be pragmatic—having to do with their aims or purposes. But if so, the choice of fitting function must itself by pragmatic in character. We see here a connection between relativism and pragmatism that ought to be obvious to people familiar with Peirce’s discussion of methods for fixing belief.

2.3 Statistical Efficiency

Let T_1 and T_2 be estimators for a parameter θ. In general, T_1 and T_2 will have different variances with respect to their estimates of θ. An estimator with smaller variance requires fewer observations to reach a given level of certainty. Hence, an appealing idea is to choose the estimator that minimizes the variance in its estimates of θ.

Initially, efficiency might look like a good candidate for an epistemic reason to choose one fitting function over another. However, there are three reasons to think that efficiency is a pragmatic reason for choosing a fitting function. First (and most importantly), evaluations of relative statistical efficiency depend on one’s choice of loss function. In the standard case, we adopt squared error loss and get the mean squared error as the criterion for judging efficiency. But such a choice is in no way forced. By adopting a different loss function, we could get a different criterion for judging efficiency. If so, then considering the relative efficiencies of some fitting functions provides a pragmatic reason for one’s choice.

Second, one may rationally choose the more efficient of two consistent estimators only if one does so on the basis of considerations of time, money, energy, or the like. If one chooses an option on the basis of considerations of time, money, energy, or the like, then one’s choice is pragmatic, rather than epistemic. Hence, if one’s choice of the more efficient of two consistent estimators is rational, then one’s choice is pragmatic, rather than epistemic.

Third, pairs of estimators (like the mean and median) have different relative efficiencies with respect to different distributions for the parameter being estimated. For example, the relative efficiency of the mean to the median is about 1.57 for the normal distribution, but it is 0 for the Cauchy distribution. As a result, efficiency opens itself up to practical trade-offs between the size of the sample one wants to draw and the sensitivity one wants to have to one's modeling assumptions. But sensitivity to one's modeling assumptions is a kind of risk that one takes. Thus, efficiency comes down to a choice between two costs: the cost of drawing a sample and the cost of being wrong about one's modeling assumptions. Hence, efficiency is a pragmatic reason for selecting a fitting function.

2.4 Robustness

Since the distribution of a parameter affects the relative efficiency of estimators for it, researchers sometimes choose a fitting function so that it will not be too sensitive to violations of the modeling assumptions. An estimator that has good relative efficiency over a wide class of modeling assumptions is *robust*. Modelers sometimes appeal to the robustness of an estimator in order to justify selecting it over its competitors. As I have argued above, efficiency is a pragmatic consideration. Since the robustness of an estimator is deeply tied to its efficiency, one ought to think that considerations of robustness are also pragmatic.

In addition, there are two further reasons to think that robustness provides pragmatic (and not epistemic) reasons for choosing a fitting function. First, there is a circularity challenge. No estimator dominates every other estimator with respect to every set of modeling assumptions. So, we are going to have to do some sort of aggregating or weighted averaging. But that just raises again the question we started with: How do we measure the goodness of fit for some estimator? That is, we could take the estimator that has the best *median* fit or the best *mean* fit or is closest in absolute value to the truth most often or is farthest in absolute value from the truth least often and so on. The reasons that one might bring to bear here again appear to be pragmatic. They are considerations about computational tractability, loss, and efficiency.

Second, the agent is still faced with a choice. For example, suppose the agent believes that the parameter has a normal distribution. The agent needs to choose between an estimator that is maximally efficient if her belief is correct but likely to be inefficient if her belief is incorrect and an estimator that is less efficient if her belief is correct but likely to still be adequately efficient if her belief is incorrect. One might think that this is a purely epistemic trade-off and hence does not show that there is any pragmatic encroachment. But an argument may be given here that is parasitic on examples used by Fantl and McGrath (2009) to support the claim that knowledge is pragmatically encroached in everyday cases. Imagine two agents, Sam and Betty. Betty wants to get an estimate that is no more than a specific distance from the truth. But she doesn't care how precise the estimate is as long as it is within that tolerance. Hence, she selects an estimator that is very robust. Sam wants to have a good chance of getting a maximally precise estimate and is happy to tolerate being wildly wrong in a few cases. Perhaps he is willing to take the risk of being wildly inaccurate, since he thinks that being wrong will not be too serious. Hence, he takes an estimator that he judges most likely to be maximally efficient, even though it is fragile.

3. No Epistemic Reasons

In Section 2, I motivated premiss [L1] of my main argument by looking at four reasons that statisticians and econometricians sometimes give for choosing a specific measure for goodness of fit. Of course, researchers might have reasons that I haven't considered, and for all I've said so far, there might be some non-pragmatic reasons for choosing a measure of fit that *no one* has considered.[5] Hence, my discussion in Section 2 should be regarded as providing a weak abductive argument designed to make [L1] worth investigating further. In this section, I defend the following principled argument for [L1].

[S1] A measure of fit defines what it is for a model or estimate to be accurate.

[S2] If [S1], then choosing a measure of fit cannot legitimately be guided by the aim of maximizing accuracy.

[S3] If there are any epistemic reasons for choosing a measure of fit, then choosing a measure of fit can legitimately be guided by the aim of maximizing accuracy.

[S4] If there are no epistemic reasons for choosing a measure of fit, then every plausible reason for choosing a measure of fit is pragmatic.

[L1] Every plausible reason for choosing a measure of fit is pragmatic.

Let me now give a gloss on the four premisses in the argument for [L1]. The first premiss is definitional (or nearly so). It says that accuracy is cashed out in terms of a measure of goodness of fit. The second premiss assumes that in order to be guided by the aim of maximizing accuracy, one needs to have some idea of what accuracy is. Either one has an account of accuracy in hand or one does not. If one has an account of accuracy, then using that account to guide one's choice

[5] Readers with a background in formal epistemology might think that philosophers in the "accuracy-first" tradition have provided a purely epistemic solution to our problem. After all, Joyce (1998) purports to give a non-pragmatic argument for probabilism. See also Leitgeb & Pettigrew (2010a, 2010b), Easwaran & Fitelson (2012), Moss (2011), and van Enk (2014). I have three reasons for thinking that the accuracy-first approach in formal epistemology does not provide a purely epistemic solution to our problem. First, some measures of fit that are plausibly ruled out in the context of defenses of probabilism cannot be ruled out in the same way in the context of our problem of goodness of fit in general statistical procedures. For example, the absolute error approach is ruled out in the accuracy-first tradition because an absolute error scoring rule is *improper*. But the analogous claim that a fitting function must be proper in order to be admissible seems entirely out of place in the setting of general measures of the goodness of fit of a model to data. Second, as Levinstein (2012) points out, plausible constraints on scoring rules admit competitors to the usual Brier score (itself equivalent to the square error approach). Third, there seem to me to be excellent arguments that the accuracy-first tradition cannot provide purely epistemic criteria for belief as it aims to do (see Levinstein 2019 and Mayo-Wilson & Wheeler 2019).

of a measure of accuracy is viciously circular. If one does not have an account of accuracy, then one cannot be guided by accuracy considerations. Either way, choosing a measure of accuracy cannot be *legitimately* guided by the aim of maximizing accuracy.

The third premiss assumes that there is an intimate connection between accuracy and epistemic reasons. The thought here is that epistemic reasons are primarily about getting at the truth, and getting at the truth makes sense if and only if one's actions can be guided by considerations of accuracy. Hence, if it isn't possible for one's choice of a measure of fit to be guided by the aim of maximizing accuracy, then there aren't *any* purely epistemic reasons one could call on to justify one's choice. The fourth premiss assumes that our reasons are either epistemic or pragmatic.[6]

4. Pragmatic Encroachment is Unavoidable in Estimation and Model Selection

In Sections 2 and 3, I argued that every plausible reason for choosing a measure of fit is pragmatic. In this section, I want to defend the second premiss in my main argument, which says that if every plausible reason for choosing a measure of fit is pragmatic, then statistical procedures used across the sciences are pragmatically encroached.

The main thought here is that every instance of estimation, testing, and model selection requires commitment to some specific measure of fit. If every reason one might give for choosing *that* measure, specifically, is pragmatic, then every instance of estimation, testing, and model selection carries with it some pragmatic commitment. And if so, then there will always be some potential change to one's pragmatic commitments that would make a difference to the

[6] Perhaps this counts too many reasons as "pragmatic." One might worry that aesthetic or ethical considerations can provide us with reasons that are neither epistemic nor pragmatic. But I don't think that people who want to cleanly separate the epistemic from the pragmatic in order to defend the epistemic purity of the sciences are going to be comforted by learning that they can call on aesthetic or ethical considerations to justify their choice of fitting function. For an epistemic purist to reject [S4] would be straining out a gnat and swallowing a camel.

estimate one ought to endorse, to the result of one's test, or to the model one ought to select in a given case. By extension, there will always be some changes in one's pragmatic commitments that would make a given estimate (held constant) or a given test (held constant) or a given model (held constant) change its status (as correct or incorrect) or its verdict (as reject or not). If so, then differences in pragmatic commitments sometimes ground differences in whether an agent knows that something is the case, which is what is meant by "pragmatic encroachment." Hence, statistical procedures used across the sciences are pragmatically encroached. What we know ultimately depends on the aims or purposes that lead to one's choice of a measure of fit.

One might object that in some cases, different measures of fit lead to identical or nearly identical estimates, test results, or models. Hence, one should commit only to results that are recommended by procedures on every measure of fit. In practical cases where one has already significantly restricted the number of competing measures of fit, the thought here is sound. But in the general case, there are no estimates or models recommended by *literally every* measure. Hence, if one declines to endorse any estimate or any model unless it is recommended by every measure, one will never endorse any estimate or any model at all.

5. Science is Pragmatically Encroached

So far, I have argued that statistical procedures used across the sciences are pragmatically encroached. In this section, I argue that consequently, science is pragmatically encroached. Science rides on the back of estimation, testing, and model selection. Hence, science cannot get very far without pragmatic commitments. Here is the argument:

[S5] If statistical procedures used across the sciences are pragmatically encroached, then tools essential to the development of scientific knowledge are pragmatically encroached.

[S6] If tools essential to the development of scientific knowledge are pragmatically encroached, then science is pragmatically encroached.

[L3] If statistical procedures used across the sciences are pragmatically encroached, then science is pragmatically encroached.

The thought behind [S5] is that science as we know it would not be possible without statistical tools that rely on measures of fit. Consider simple curve-fitting. Glymour (1980, 322) is simply poetic in his account of the historical importance of curve-fitting:

> A GREAT DEAL of the most basic scientific inference consists of the following sort of thing: an observer measures or calculates values for two (or more) quantities, and obtains a set of paired values; from these paired values he infers a functional relation between the quantities measured. Inferences of this kind can be found in perhaps the majority of experimental reports published in the physical sciences, and some of them constitute great steps in the history of science: Kepler's laws, Coulomb's law, the law of specific heats, various radiation laws, Hubble's law, the gas laws, and so on; the list of important relations argued for in this way is nearly endless.

Of course, the oldest historical cases were not statistically sophisticated. But as datasets have grown both in terms of the number of statistical units observed and in terms of the number of variables measured, guesses constrained merely by *implicit* measures of fit stop being adequate for the advancement of science. Moreover, as Spanos (2007, 1059-1064) argues, implicit statistical modeling assumptions, including assumptions about proper measures of fit, are crucial to the vindication of even ancient curve-fitting enterprises, for it is precisely failures in the *statistical* modeling that vindicate Kepler’s model of the orbit of Mars as compared with Ptolemy’s.

While it might be possible to have some kind of science without any estimation, theory testing, or model selection, it is hard to see how such a science could grow and develop in anything approaching the way in which our own science has grown and developed. If one were to excise estimation, testing, and model selection from the scientific enterprise as we have it today, the residue would not be recognizable to us as science. I am not attempting here to provide a demarcation criterion: I am only arguing for the necessity for science of broadly statistical procedures that make use of some measure of fit (perhaps implicitly). I am not arguing that the use of estimation, testing, and model selection are *sufficient* to make an inquiry scientific—though I am inclined to think that having more statistical and mathematical modeling is an *indicator* that an investigation is scientific (cf. Pigliucci 2013). Nor am I arguing that every activity in the scientific enterprise consists in estimation, testing, or model selection. The construction of scientific instruments, the collection of data, the invention of categorization schemes, and the proposal of structural accounts of natural phenomena are all examples of activities that are properly scientific in character but do not consist in statistical procedures.

I think [S6] is obvious, which I suppose is dangerous to admit. But it seems to me that if something essential to the growth and development of science as we find it is pragmatically encroached, then science itself is pragmatically encroached. A little yeast works through the whole dough. Summing up, then, I have defended all of the premisses in my main argument. Taken altogether, they deliver the conclusion that science is pragmatically encroached.

6. Concluding Remarks on Inductive Risk

Douglas (2000) argues that "because of inductive risk, or the risk of error, non-epistemic values are required in science wherever non-epistemic consequences of error should be considered" (559). In concluding, I want to make some brief remarks on the relationship as I see it between my arguments for pragmatic encroachment by way of measures of goodness of fit and Douglas' argument against the value-free ideal in science. To begin with, I agree with Douglas in thinking that the value-free ideal cannot be maintained in science. And I think that examples like hers could be used to good effect in illustrating points I have tried to make regarding statistical efficiency and robustness in Sections 2.3 and 2.4. But on my reading, Douglas is too modest in the degree to which science is infiltrated by considerations of practical value. She seems to think that non-epistemic virtues are only required in cases where practical decisions need to be made and where risk of error has real-world consequences. Although Douglas does not mention Neyman, her view of statistical testing seems to be essentially Neyman's (1957) "inductive behavior." A proponent of the value-free ideal might hold out hope that a Fisherian approach embracing full-blooded scientific *inference* and not just decision-making might deflect or evade the force of Douglas' arguments. But if I am right, challenges to the value-free ideal do not depend on Neyman's account of statistical testing or anything similar. It seems to me that non-epistemic values go deeper than one might suppose on the basis of Douglas' arguments, and I hope that my arguments will be understood as cutting deep enough to expose the pragmatist bones holding up the body of scientific knowledge.

References

Birkes, D., and Dodge, Y. (2011). *Alternative Methods of Regression*. John Wiley & Sons.

Douglas, H. (2000). Inductive Risk and Values in Science. *Philosophy of Science* 67(4), 559-579.

Easwaran, K., & Fitelson, B. (2015). Accuracy, coherence, and evidence. *Oxford studies in epistemology*, *5*, 61-96.

Fantl, J., & McGrath, M. (2009). *Knowledge in an uncertain world*. Oxford University Press.

Farebrother, R. W. (1999). *Fitting linear relationships: A history of the calculus of observations 1750-1900*. Springer Science & Business Media.

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*(1), 1-35.

Glymour, C. (1980). *Theory and Evidence*. Princeton University Press.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of science*, *65*(4), 575-603.

Leitgeb, H., & Pettigrew, R. (2010a). An objective justification of Bayesianism I: Measuring inaccuracy. *Philosophy of Science*, *77*(2), 201-235.

Leitgeb, H., & Pettigrew, R. (2010b). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science*, *77*(2), 236-272.

Levinstein, B. A. (2012). Leitgeb and Pettigrew on accuracy and updating. *Philosophy of Science*, *79*(3), 413-424.

Levinstein, B. A. (2019). An objection of varying importance to epistemic utility theory. *Philosophical Studies*, 176(11), 2919-2931.

Mayo-Wilson, C., and Wheeler, G. (2019). Epistemic Decision Theory's Reckoning. http://philsci-archive.pitt.edu/16374/1/25a_EDTR.pdf

Moss, S. (2011). Scoring rules and epistemic compromise. *Mind*, *120*(480), 1053-1069.

Neyman, J. (1957). “Inductive Behavior” as a Basic Concept of Philosophy of Science. *Revue de l'Institut International de Statistique*, 7-22.

Pigliucci, M. (2013). The demarcation problem. A (belated) response to Laudan. *Philosophy of pseudoscience: Reconsidering the demarcation problem*, 9.

Portnoy, S., and Koenker, R. (1997). The Gaussian Hare and the Laplacian Tortoise. *Statistical Science* 12(4), 279-300.

Prescott, P. (1975). An Approximate Test for Outliers in Linear Models. *Technometrics* 17(1), 129-132.

Spanos, A. (2007). Curve Fitting, the Reliability of Inductive Inference, and the Error-Statistical Approach. *Philosophy of Science* 74, 1046-1066.

Stigler, S. (1977). Do Robust Estimators Work with *Real* Data? *The Annals of Statistics* 5(6), 1055-1098.

van Enk, S. (2014). Bayesian Measures of Confirmation from Scoring Rules. *Philosophy of Science* 81(1), 101-113.