Received 28 May 2024
Revised 12 June 2024
13 June 2024
Accepted 24 June 2024

Using predictive methods to assess observation and measure importance

William M. Briggs
Statistician to the Stars!, Charlevoix, Michigan, USA

Abstract

Purpose – This study aims to find suitable replacements for hypothesis testing and variable-importance measures.

Design/methodology/approach – This study explores under-used predictive methods.

Findings – The study's hypothesis testing can and should be replaced by predictive methods. It is the only way to know if models have any value.

Originality/value – This is the first time predictive methods have been used to demonstrate measure and variable importance. Hypothesis testing can never prove the goodness of models. Only predictive methods can.

Keywords Bayes factors, Hypothesis testing, Predictive modeling, Replication crisis

Paper type Research paper

1. Introduction

There are myriad arguments against *p*-values, Bayes factors and all kinds of so-called hypothesis testing. The discussion is endless, and the argument is going in circles. I have no wish to burden the reader with yet another review but a sampling of pertinent literature (Wasserstein and Lazar, 2016; Berger and Selke, 1987; William and Hung, 2019; Briggs *et al.*, 2019; Colquhoun, 2014; Goodman, 2001; Greenland *et al.*, 2016; Harrell, 2017; Nguyen, 2016; Trafimow *et al.*, 2018). The literature on this topic is huge and so well known that we spend no time covering this well-trodden ground except for one facet, which I believe is underappreciated.

Perhaps the best argument against these parameter-centric or hypothesis-testing practices is this: that if they work, they should always work; if we know that there are instances in which they do not work, we are right to suspect they never do, and if we suspect they never do, then they should not be used. This argument will be given briefly in the next section. Whether or not this argument is convincing to the reader, the remainder of the paper still has use in demonstrating usefulness of predictive methods.

The rest of this paper is devoted to showing how predictive methods, which may be considered whole-model methods, can work to show the importance of observations and measures that are part of models. I use the term "whole-model" to mean a model as given in its complete form – it's the form of use in the real world, which is to say, its predictive form. The philosophy of predictive methods is given in Briggs (2016) and Geisser (1993), among

JEL Classification — C4, C5

Erratum: It has come to the attention of the publisher that the article, Briggs, W.M. (2024), "Using predictive methods to assess observation and measure importance", *Asian Journal of Economics and Banking*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/AJEB-05-2024-0066 failed to include the JEL codes, these have now been added. The publisher sincerely apologises for this error and for any inconvenience caused.

Asian Journal of Economics and Banking
Vol. 8 No. 3, 2024
pp. 354-365
Emerald Publishing Limited
e-ISSN: 2633-7991
p-ISSN: 2615-9821
DOI 10.1108/AJEB-05-2024-0066

others. Methods that are centered on the "model guts," such as values of unobservable parameters, can never be wholly convincing, at least because it is always possible to find a model that fits any set of data arbitrarily well.

Here we demonstrate two simple procedures: checking the importance of individual observations to a model's predictive ability and the importance of individual measures (or variables) to that predictive ability. This is work similar to that in Meyer *et al.* (2002) and Geisser and Eddy (1979), but with a different emphasis on philosophy.

All figures, words, grammar, punctuation and wit are by the author.

2. An argument against hypothesis tests

This section simplifies the argument first given in Briggs (2016). A gambler sits at an American roulette wheel and is tracking the results. An American wheel has 18 red slots, 18 black and 2 green slots. Given these premises, we deduce that the probability that the wheel comes up red is 18/38. The wheel has come up red the past 10 of 12 times. The gambler is a fan of hypothesis testing and calculates a binomial test against the chance of 18/38. This gives a p-value of 0.018.

The gambler, who recalls his frequentist theory, therefore concludes that black is "due." Or perhaps he is also argumentative, and he concludes the wheel is imbalanced and accuses the casino of cheating.

There is not a statistician alive who claims to follow frequentist theory and will not back that gambler to the last drop of his blood. Frequentist theory insists that if the p-value is small, the "null" hypothesis *must* be rejected.

Now I have never met a person who claims to follow frequentist theory and does so in real life. Not consistently. They always act at least like a Bayesian, especially when it comes to interpreting confidence intervals on parameters. Never once have I seen a statistician, in writing outside of introducing the topic in textbooks, act like he believes the theory that he teaches. For instance, all that can be said about a confidence interval is that the "true" value of the parameter is in the interval or it isn't. Or that if the procedure (whatever that might be) that gave rise to the data that led to calculating the confidence interval were repeated an *infinite* – not large: infinite – number of times, then such-and-such percent of those intervals would cover the "true" value.

This correct interpretation is absurd because it says nothing about the interval at hand. Which is why all working statisticians ignore it and adapt a Bayesian interpretation of the confidence interval, which, of course, invalidates the theory. It should not be used.

People who claim to believe in frequentism do the same thing with p-values. There is no guide, whatsoever, inside the theory that says *this* time the p-value can be trusted and that time it cannot, so ignore it. The theory just insists that p-values work; therefore, they must work all the time.

But, of course, no statistician believes that. They violate the theory with wild abandon, saying, as they should say here to the gambler, that this p-value cannot be trusted, but that one can. In other words, they act as Bayesians, sneaking in prior information on the parameters that are forbidden – outright outlawed – by frequentist theory. They allow themselves to cheat like this because they don't formally write down these priors or make math of them, and so they can pretend that if they can't see it on paper, then they don't exist.

This cheating, this ignoring of the strict demands of frequentist theory, occurs constantly. So much so that it is far past time to lend any support to frequentist theory. At the very least, Bayesian theory should take its place.

The problem with Bayes, though, is that it is still parameter-centric in nearly all applications. Bayes factors are computed against unobservable parameters. There is no way to test these parameters because they cannot be observed. Even if the parameters are correctly assigned, in unique true and not *ad hoc* models. There is no way to know that unless the model itself is tested. And the only way to test the model is against reality.

Hence, the argument is that models should be tested against reality. And the only way to do that is not to just measure how well models fit the data used to build them, which is what hypothesis testing does, because this is so easy a test it can nearly always be passed, especially with the freedom people have to tweak models to better fit data.

3. The simplest form

All models fit the schema:

$$p(y|x), \tag{1}$$

where x is the *entire* list of assumptions, premises, evidence, observations, theory and whatever else is used, including the long list of implicit premises, which are not written down but are always there. The y is the object we wish to say something about, based on the particular value of x that is specified. The "dimensions" of y and x are given by the model too. Included in x are those assumptions, etc. That give us the form of $p()$. The function $p()$ either specifies an exact value for y or gives a probability (distribution, etc.) or some other form of uncertainty (a $\pm$, perhaps). This is almost always an *ad hoc* form in statistical models.

Causal models specify values, and correlational models give probabilities. This should be obvious. If we *knew* the *full* cause of y, it would be in x, and therefore, we could say exact things about y. If we do not know the full cause, the best we can do is correlation, which may, of course, include *some* but not *all* causal elements.

This schema fits every model, from simple coin flips to quantum mechanics to cosmology to the most "powerful" artificial intelligence (AI) to global ocean-atmosphere-coupled climate models.

The predictive method could not be easier. It is this: (1) specify or wait for or make x occur in the world, (2) see what happens to $p(y|x)$ and (3) then compare the outcome with what happened to Y|X in the world. Y itself is a proposition; it is a statement about what happened, or rather, what was observed, in the world. (By *world,* I mean everything there is.)

Perhaps the simplicity of this form is what makes it appear unusual, but not in every field. Some disciplines, like the aforementioned AI and areas like engineering, make use of it exclusively; see, e.g. Romano *et al.* (2019). Others find comfort in "hypothesis testing," which we have seen is error-prone at best and only says things about the x (e.g. parameters live inside x). As we shall see, the more widespread adoption of the predictive method may be delayed by the difficulty that one must specify some kind of decision rule that measures the "distance" between Y|X and $p(y|x)$; i.e. we must pick some $d(p(y|x), \text{Y|X})$, which is problem- and decision-dependent. As we'll see, this beats the entirely *ad hoc* and groundless approach of testing.

Quite obviously, $d()$ is problem-dependent, even if $p(y|x)$ itself is not. A useful or valuable model in one context can be entirely useless or harmful in another. This means two decision-makers might not agree on which $d()$ is best. There may be no one best measure between predictions and the world. Even models that have a rotten agreement with the world can still be valuable to their creators if those models are considered important in other ways.

The real and obvious benefit of the predictive method is that a model is built using whatever means are preferred, means that can even remain opaque, mysterious or secret. If a modeler wants his model to be evaluated, all he needs to do is issue $p(y|x')$ for whatever value of x' he thinks is useful to decision-makers, e.g. he releases x' but believes x; x' may be a subset of x. The x' contains at least enough information so that the model can be tested against reality. This may be somewhat difficult to understand until we have seen an example.

4 . The practical form

Another way to write the model schema is this:

$$p(y|x) \equiv p(y|ewo), \tag{2}$$

where $x \equiv ewo$, in which the complex proposition x has been broken up into more manageable pieces, here labeled o for past observations considered probative of y, w is the way the world is now or the way it is assumed to be or will be and e is all the other premises, explicit and tacit, used in assuming the model form; the explanation of the world, if you will. This is helpful because often e stays roughly the same when people are thinking up models. The model form is either directly specified or deduced from e. If the model is open, all details of how the model is created are in e. If the model is (for whatever reason) secretive, only those elements that specify the form are given. For instance, we are told e' = "The model is a logistic regression," but given no other justification for this except its declaration (this is most models in statistics). It's also useful to separate, in notation, old from new (as it were) observations or measures. In case it is not obvious, w and o are the same; i.e. they must have the same elements but possibly different values.

The observations o are of the form $o = o_{ij}$, where i represents the ith measure and j the instance of the ith measure. This means that w has the same form, $w = w_{ik}$, where the measures i must match $o_{i\cdot} = w_{i\cdot}$. In traditional parlance, one says that the "variables" are the same, but their values in w in new suppositions, predictions, or scenarios might not be equal to the old observations. This will become clearer in a moment.

4.1 Importance of past observations

We would like to learn how important any given observation is to the model. It would seem that there are two meanings of importance for past observations: of (or in) the model and of the model's predictions.

Importance in the model might mean how much a given observation affects the formation of the model. It could work like this: (1) derive (or fit, or whatever language you like) the model using all information; (2) re-derive the model n times, each time leaving out $o_{\cdot j}, j = 1, 2, \ldots, n$; (3) compare the change in the model for each removed observation. This comparison can be done with the full model, which uses all n observations, or between the smaller "$-j$" models.

This requires a measure of "distance" or importance. Which is best, of course, depends on the decision being made with the model and its measures. There is *no* universal "best" distance or metric of similarity. And one should never be advocated.

Yet here is where the problem with this view arises. How does one judge the difference or distance between $p(y|ewo_j)$ and $p(y|ewo_{-j})$? It cannot be done without specifying a w. This being so, one idea is to abandon all thoughts of the final model by removing w and then measure how well the unobservable parameters relate o to itself, given $p()$, which is to say, given the evidence e used to deduce the model form. This is the realm of hypothesis testing, either with p-values, confidence intervals on the parameters, or using Bayes factors. It is to put the primacy of model fit over model usefulness or truth.

The problem, of course, is that no matter how close some measure of $p(y|eo)$ is to o itself, given the model or its guts (like parameters), it tells us nothing about how good $p(y|ewo)$ is. That is, the model may be said to "fit" o well, even perfectly, but this does not translate into the model making good predictions of w. And if the model cannot make good predictions of w, then there is no reason to believe the model (as deduced from the theory e) is any good at *explaining* the world either.

This means there is no way to directly check the model or the importance of any single observation, except as predictions. Thus, we have to examine observational importance predictively. This is easy to do.

The idea is simple and is similar to many predictive cross-validation-type methods, see, e.g. Lu and Ishwaran (2017). Create $p_j(y|ewo_{-j})$ for each observation $j = 1, 2, \ldots, n$, i.e. "fit" the model minus observation j, specify a w, make the prediction and compare $p_j(y|ewo_{-j})$ with the resulting Y|W; if w can be controlled (in the true experimental sense of control and not the

weak and misleading meaning of statistical control). If w cannot be controlled but it and y can be observed, an equivalent method is to measure Y and discover which values of w held when Y was measured.

To do this, we need the distance or importance score $d_j(p(y|ewo_{-j}), \mathrm{Y}|\mathrm{W})$. For the Y| W pair, we have d_j, $j = 1, 2, \ldots, n$. We will here suppose smaller d are better (but this is of course entirely problem-dependent). These d_j can be ordered from smallest to largest, and this describes their importance. Observations that do not move d_j from d (the distance of the model for all observations used) imply the observation is not as important as those that do move d_j from d. Importance can be beneficial or harmful.

If one has a collection of Y|W, instead of just one new observation, then the average of d_j can be used. Examples will make this clear. Or any other problem-dependent function instead of an average. Whatever distance would be used in real life by the entity relying on the model should be picked. Again, a model useful to one person can be useless, or even harmful, to another.

4.2 Importance of measures

This idea here is the same as the importance of individual observations. The model is derived with all measures, and then once again, each time leaving out a measure, for $i = 1, 2, \cdots q$ all q measures. Again, a w is specified or observed, along with the resulting Y, and the model is compared against reality. The distance $d_i(p(y|ew_{-i}o_{-i}), \mathrm{Y}|\mathrm{W})$ is computed for each i. These distances can again be ordered, which gives the idea of the importance of each measure in the same way. The d_i can also be compared against d of the full model, i.e. the one with all measures in it. Once again, the examples will make this clear.

4.3 What about w*?*

As said, the ideal case is one in which, after the model is in hand, w can be experimentally manipulated, which is to say, controlled. Less ideal, but still equally valid, is when w can be observed. It is less ideal because, when observed, the temptation to move from correlation to claims of cause is sometimes too much for some to resist, as the scientific literature amply demonstrates.

But what to do if there no w at hand? That is, all that is available are the old observations o, with new ones (the w) not in sight or too expensive or time consuming to gather. What is best, of course, is to wait until such w do become available. There is no other way to test the validity of the model except by comparing it against reality independently of the model-building process. If we have to wait, if or it is costly, that is the price that must be paid. Alas, it rarely is paid.

As said, the old way was to simply announce the model fit either in the form of hypothesis tests or Bayes factors or by giving information on the unobservable parameters inside the model. These practices won't do, not if used to justify the models.

Still, it is true that, given that we have seen o before, it seems possible we can see observations *like* o again. This is not unreasonable. And so the practice of pretending o is w has developed, which is to say, we treat the old observations as if they were new ones. We let $w = o$, which means we must also let the old y become the "new" Y|W.

The problem with this is obvious, or should be. It's because the models are fit to match o and y closely. That doesn't mean that the model is wrong, but that, as I keep insisting, because the model fit is only a necessary but far from sufficient criterion to demonstrate model validity. Real proof of a model's goodness is still lacking.

Because the model has been fit to o and y, if we pretend $w = o$, we are necessarily painting for ourselves a pretty, over-optimistic picture. If we have some procedure, algorithm, or computation that uses the same observations o that go into fitting the model to judge the model, we almost certainly will be over-certain if we claim that the model well represents reality.

A variant of this is some form of "cross-validation." The observations are split in some way, $o = o_t o_v$, where o_t is the "test" set, i.e. the observations that are used to fit the model, and o_v is the validation set, i.e. $w = o_v$ (and the same split for the observed y). This works as the predictive method should, except for two glaring problems:

(1) After the cross-validation is complete, the model is re-fitted, not to o_t but to the "full" o, i.e. to $o = o_t o_v$ and these two models are *not* the same! That is, $p(y|ewo_t) \neq p(ewo_t o_v)$. However well $p(y|ew_v o_t)$ predicts $Y|W=O_v$, it does not mean that $p(ewo_t o_v)$ will predict some new Y|W equally well, or at all well.

This follows from the insistence that *any* change in the conditions (after the bar in the model) logically implies a *different* model. And all models must be judged independently. This is rarely recognized.

(2) After $p(y|ew_v o_t)$ is compared against $Y|W_v = O_v$ and the model's strengths and weaknesses are revealed, the temptation to change e is almost never avoided. For instance, some measures will be left out of o (and thus out of w), or some will be changed or modified, or perhaps new ones will join. Or maybe new observations are added or subtracted. Or the model form itself is change ("Let's use a gamma instead of a normal error."). But it is the case this new model is *not* the same as the old, i.e. $p(y|ewo) \neq p(y|e'w'o')$ and we do not have proof of the new model's ($p(y|e'w'o')$) goodness.

The only way to test a model is to put it against the real world in a "replication." The way engineers build new aircraft engines, for instance.

Still, with that very great caution in mind, it can still be useful to suppose $w = o$, as long as the very strict limitations we have outlined are kept in mind and nothing great about the model itself is claimed.

5. Example

We will use ordinary, one-dimensional regression as an example, since almost all readers will be familiar with it; it is a ubiquitous technique. To further simplify matters, we'll use logistic regression, which gives a single probability to an observable as a function of some number of measures. There is nothing at all special about this model except for its simplicity. Keep in mind that the techniques developed here can be used for *any* model, no matter its sophistication or form of y.

The dataset is from UCLA (n.d.) and describes whether or not students were admitted to a graduate program, given their undergraduate grade point average (GPA), graduate record examination (GRE) score and a subjective rank, 1 (best) through 4 (worst), of their undergraduate program. One might expect that higher values of all these would lead to a greater frequency of being accepted. And that's what we find by a cursory examination of the observations, as given in Table 1.

Rank	1	2	3	4
	0.54	0.36	0.23	0.18
GRE	(220, 520)	(520, 580)	(580, 660)	(660, 800]
	0.23	0.26	0.39	0.40
GPA	(2.26, 3.13)	(3.13, 3.4)	(3.4, 3.67)	(3.67, 4)
	0.21	0.24	0.40	0.42

Note(s): Except for rank, which is presented as is, the other measures are split by quartiles. As expected, better (higher) values of the measures are associated with greater frequencies of Admittance
Source(s): Author

Table 1. A rough cut of the observations for each measure and the frequency of being admitted

We cannot here build a causal model because we cannot know the causes of admittance in each of the ($n = 400$) cases. It is anyway clear from the data itself that a fixed algorithm for admittance was not used. Such an algorithm, were it present, might have been of the form, e.g. "If rank ≥2, GRE ≥700 and GPA ≥3.5, Admit." So our model is correlational.

We use the model

$$\Pr(\text{Admit}|ewo), \tag{3}$$

based on an ordinary logistic regression (this is our e), with o being the terms mentioned, all entered in the model additively (that additivity is also part of e). There is no special reason for additivity; the terms could have just as well been put in multiplicatively, or in some combination, say rank additively and GPA times GRE. Again, the terms are entered additively, and the form of the model itself is our e.

Also included in e are the propositions that put parameters to each of the measures o and whatever evidence is used to specify the priors on those parameters. Here we use the familiar parameters of a linear model, along with the default priors on these parameters as given in the RSTANARM package (version 2.21.3) in R. We also used the setting "iter = 10,000" to give the algorithm more time to run.

We stress the setting of these "dials", as they were are also part of e, which means that if we change *any* part of e, including all the "random" numbers that are used in the Markov chain Monte Carlo (MCMC) algorithm, we *necessarily* change the model into a new model. That means, and I absolutely insist on this, that if the MCMC algorithm is run twice, assuming the random seed has changed, the end result is two *separate* models. This will rub those who believe there are "true" values of the parameters that are somehow conjured into existence once the model form is specified, which means there must be "true" values for every possible model form anybody can think of with this set of data, which is a lot. The predictive view does not hold with this and only asks us to evaluate the evidence at hand, codified in the assumptions on the right-hand side of the bar.

As written, 3 is the predictive posterior of the logistic regression, here using $w = o$ and with $y = \text{Y|W}$. That is, we assume new observations are exactly like the old ones.

The score we picked to measure the distance between the model and its predictions is the Brier score:

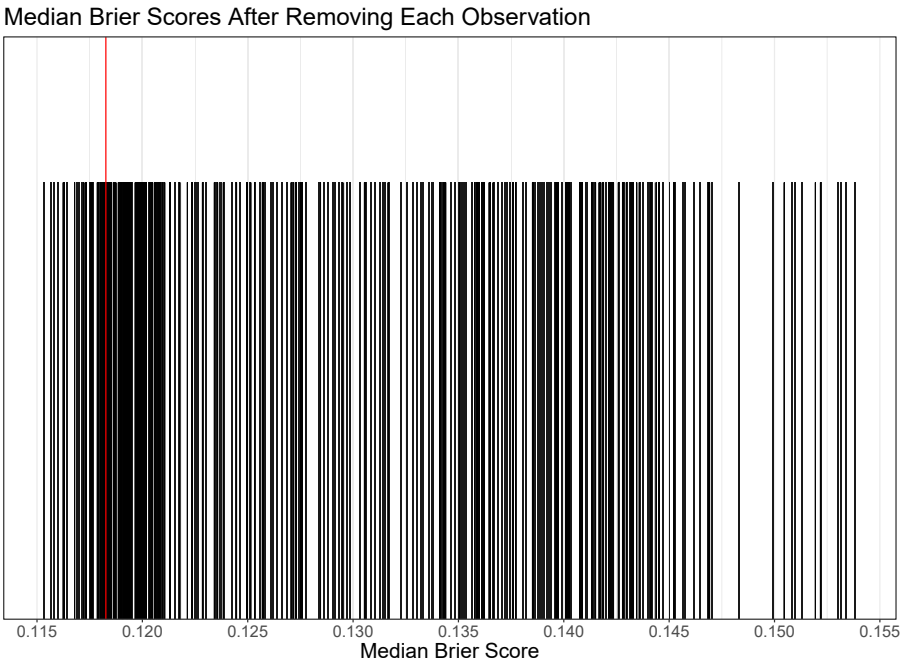
$$d(p(y|ewo), \text{Y|W}) = (\Pr(\text{Admit}|ewo) -, \text{Y|W})^2. \tag{4}$$

Now, whether this is a score useful to the reader, or to any model user, is unknown. It is picked here only because of its popularity and simplicity of form. Recalling admit can only be 1 or 0, smaller Brier scores are better. We also chose to look at the median Brier score, when setting $w = o$, i.e. all the past observations. Again, whether this is interesting or useful, or some other distance would be preferable, depends on the uses to which the model is put. These choices, however superior or suboptimal they are, at least give a flavor of how the predictive method works.

Figure 1 shows the median Brier score after removing observation j, for each $j = 1, 2, \ldots, n$ and letting $w = o$ and then predicting the probability $y = \text{Y|W}$. The vertical red line is the median Brier score for the full model, i.e. the fit using all observations. The median Brier score for the full model is 0.118.

Removing observations to the right of the red line *worsens* the Brier score, showing the observations to be positively important to the predictive *success* of the model, because when they are removed, the model gets worse. Removing observations to the left of the red line *improves* the Brier score, showing these observations to be harmfully important.

Table 2 shows the four worst and four best observations, according to whether their removal improved the median Brier score or worsened it. Recall that the median Brier score using the model fit with all observations was 0.118. Removing the four worst observations does improve this, but only very slightly, to a low of 0.115, which means these observations



Note(s): The overall median Brier score of the model fit using all data is the red vertical line at 0.118. Observations by their removal that improve the model make the scores worse than the overall, and vice versa for observations that degrade the model

Source(s): Author

Figure 1.
The median brier scores after removing each observation and letting $w = o_{-j}$

Admit	GRE	GPA	Rank	Brier
Four worst observations				
1	540	3.78	2	0.115
1	620	3.75	2	0.115
1	540	3.77	2	0.116
1	480	2.62	2	0.116
Four best observations				
0	700	3.92	2	0.153
1	560	2.98	1	0.153
1	640	3.19	4	0.153
0	400	3.08	2	0.154

Note(s): Recalling the overall median Brier score was 0.118, the scores here may be used as a comparison

Source(s): Author

Table 2.
The four worst and four best observations, according to whether their removal improved the median Brier score or worsened it

are "dragging the model down," as it were, but not by much. It is interesting the ranks are all on the "second place", and the GPAs and GREs are on the modest side, and all were admitted.

What this means is that I leave it to experts in admission policies, which is the point: it may have some meaning the statisticians know nothing about. Importance is relative and decision-dependent.

The four best observations change the median Brier score by a lot more. The values of the measures and observables (the outcome) are more varied. Is this difference of 0.118 to about 0.154 enough to be important to a decision-maker?

Well, that depends on, as always, what uses the model will be to. What we can say is that a picture like Figure 1 can be useful in studying how influential observations are and in a way that matters to decision-makers. For instance, observations that appear to be "noise" or are suspicious in some way can be easily identified in this way.

The same procedure is now done for measures, of which there are only three, with the results given in Figure 2.

Figure 2 shows the results of the procedure. The Brier score medians for the model fit using all observations are presented after removing each measure as noted. The overall model using all measures is in red, as before. Each measure improves the model, because removing any of them makes the score worse.

Take GRE as an example. Removing it reduces the median Brier score by about 0.002. Is the reduction of two thousandths in the Brier score worth including the GRE? After all, it costs something to include this data, let alone the personal costs of taking and administering the test.

That is not a question I can answer, but a college administrator of graduate studies might be able to. Model performance is neither cost free nor making any observations. Collecting GRE, or any measure, is costly and must be balanced against the predictive performance of the model.

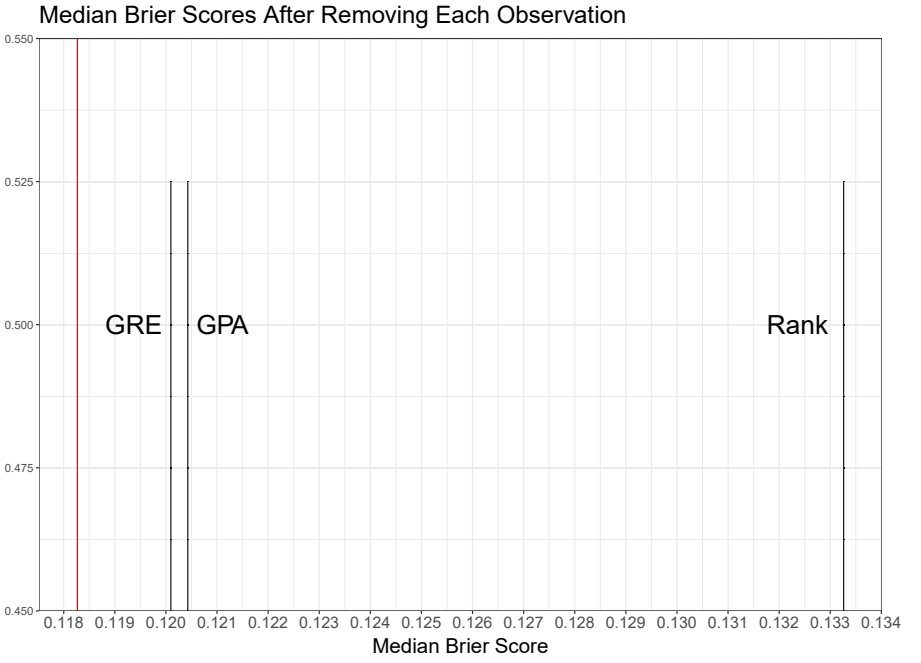


Figure 2.
The brier score medians for the model fit using all observations but after removing each measure as noted

Note(s): The overall model using all measures is in red, as before. Each measure improves the model, because removing any of them makes the score worse

Source(s): Author

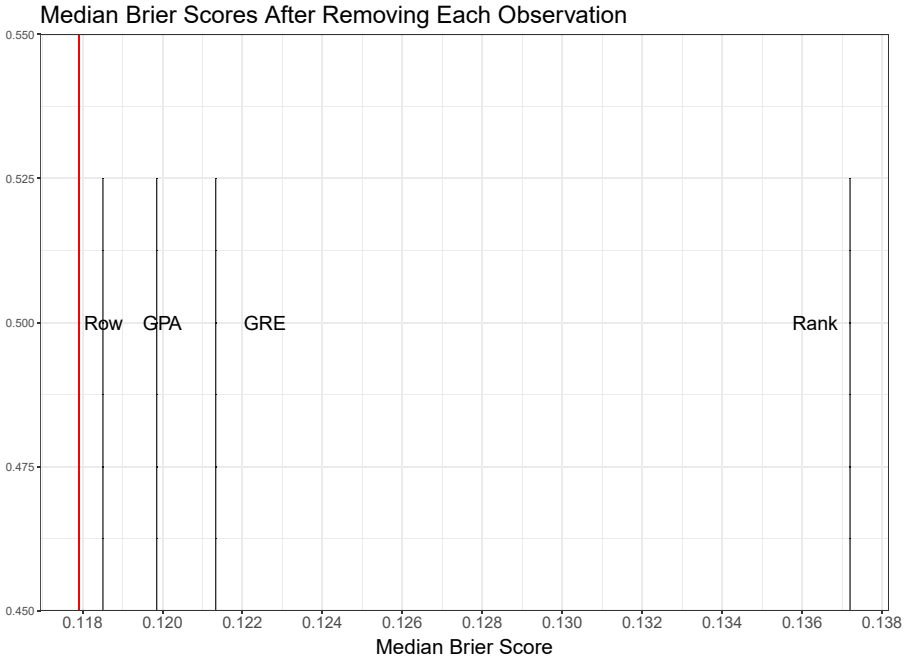
Unfortunately, for the purposes of demonstration, all measures here were helpful in improving the model, at least in the sense defined here and using this performance score. But suppose, merely for the sake of demonstration, a statistician mistook the row numbers for the data itself, not paying heed. This has indeed happened in real data, as many statisticians can tell you. Perhaps the data creator thought the row numbers were some kind of precedence, with higher numbers being better, or whatever. Let us redo the analysis with this in mind, resulting in Figure 3.

Adding a row does indeed improve the model, but only in the presence of the other measures. The improvement is minuscule. Is it so small that a decision-maker would eschew it? Well, that depends on the decision. Here, at least, we can see the row is adding almost nothing to the predictive ability of the model.

One can repeat this kind of thing, adding "noise" and seeing that, sometimes, this noise worsens the model and sometimes improves it. If one does not know it is noise and thinks the measure is possibly probative, which is the situation most of us are in most of the time when we use regression, the only way to find out is to test the model with that measure in it and see. So once again, we cannot escape the predictive method.

6. Conclusion

I wish that hypothesis testing, in frequentist or Bayesian form, would be eliminated. Even if the philosophical foundations of testing were sound, and I and others claim they are not, they do not give true indications of model soundness, usefulness or truth. They are one-size-fits-all procedures that do not fit all situations.



Note(s): The overall model using all measures is in red, as before. Each measure improves the model, even Row, because removing any of them makes the score worse

Source(s): Author

Figure 3. The brier score medians for the model fit using all observations, but after removing each measure, with "noise" added in by assuming the row numbers of the data were an actual measure

Worst of all, hypothesis-testing addresses the interiors of models, usually speaking of unobservable model parameters. That being so, there is no way to ever check in any real situation whether the hypothesis testing is giving the correct results. There is no empirical test of these parameters, because of course, they can never be measured, except in highly artificial situations that never apply to models of the world.

The examples used above only deleted one observation at a time to see the influence or importance of that observation, given all the others. And the same was true for the measures: one at a time was deleted. Of course, this may not be sensible for every problem. Groups of measures or observations may belong naturally together, and so it makes little sense to analyze their absence singularly, whereas it might make sense to measure their importance in bulk. The point here is not the exact subtractions made in the example, but that the technique of showing importance by the predictive value of observations or measures is valid and not subject to the vicissitudes of hypothesis testing.

It still remains, however, that the only good test of a model is to use it to predict independently of the process used in creating the model. Anything else leads to over-certainty.

References

Berger, J.O. and Selke, T. (1987), "Testing a point null hypothesis: the irreconcilability of p-values and evidence", *JASA*, Vol. 33 No. 397, pp. 112-122, doi: 10.2307/2289131.

Briggs, W.M. (2016), *Uncertainty: the Soul of Probability, Modeling & Statistics*, Springer, New York.

Briggs, W.M. (2019), "Everything wrong with *p*-values under one roof", in Kreinovich, V., Thach, N.N., Trung, N.D. and Thanh, D.V. (Eds), *Beyond Traditional Probabilistic Methods in Economics*, Springer, New York, pp. 22-44.

Colquhoun, D. (2014), "An investigation of the false discovery rate and the misinterpretation of *p*-values", *Royal Society Open Science*, Vol. 1 No. 3, pp. 1-16, doi: 10.1098/rsos.140216.

Geisser, S. (1993), *Predictive Inference: an Introduction*, Chapman & Hall, New York.

Geisser, S. and Eddy, W.F. (1979), "A predictive approach to model selection", *Journal of the American Statistical Association*, Vol. 74 No. 365, pp. 153-160, doi: 10.1080/01621459.1979.10481632.

Goodman, S.N. (2001), "Of *p*-values and Bayes: a modest proposal", *Epidemiology*, Vol. 12 No. 3, pp. 295-297, doi: 10.1097/00001648-200105000-00006.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. and Altman, D.G. (2016), "Statistical tests, *P*-values, confidence intervals, and power: a guide to misinterpretations", *European Journal of Epidemiology*, Vol. 31 No. 4, pp. 337-350, doi: 10.1007/s10654-016-0149-3.

Harrell, F. (2017), "A litany of problems with *p*-values", available at: https://www.fharrell.com/post/pval-litany/

Lu, M. and Ishwaran, H. (2017), "A machine learning alternative to *p*-values", *arXiv preprint arXiv: 1701.04944*.

Meyer, M.C. and Laud, P.W. (2002), "Predictive variable selection in generalized linear models", *Journal of the American Statistical Association*, Vol. 97 No. 459, pp. 859-871, doi: 10.1198/016214502388618654.

Nguyen, H.T. (2016), "On evidence measures of support for reasoning with integrated uncertainty: a lesson from the ban of p-values in statistical inference", in *Integrated Uncertainty in Knowledge Modelling and Decision Making*, Springer, pp. 3-15.

Romano, Y., Patterson, E. and Candes, E. (2019), "Conformalized quantile regression", in Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E. and Garnett, R. (Eds), *Advances in Neural Information Processing Systems*, Curran Associates, Vol. 32.

Trafimow, D., Amrhein, V., Areshenkoff, C.N., Barrera-Causil, C.J., Beh, E.J., Bilgiç, Y.K., Bono, R., Bradley, M.T., Briggs, W.M., Cepeda-Freyre, H.A. and Chaigneau, S.E. (2018), "Manipulating the alpha level cannot cure significance testing", *Frontiers in Psychology*, Vol. 9, 699, pp. 1-7, doi: 10.3389/fpsyg.2018.00699.

UCLA (n.d.), "UCLA advanced research computing: statistical methods and data analytics", available at: https://stats.idre.ucla.edu/stat/data/binary.csv (accessed 10 May 2024).

Wasserstein, R.L. and Lazar, N.A. (2016), "The ASA's statement on *p*-values: context, process, and purpose", *American Statistician*, Vol. 70, pp. 129-132.

William, M.B. and Hung, T.N. (2019), "Clarifying ASA's views on p-values in hypothesis testing", *Asian Journal of Economics and Banking*, Vol. 3, pp. 1-16.

Corresponding author
William M. Briggs can be contacted at: matt@wmbriggs.com