A connection between discrepancy function estimation and the p-value

Andrew Neath* Joseph Cavanaugh† Benjamin Riedle‡

Abstract

Consider the common statistical problem of using data to decide between a null model and a more general alternative. Within the significance testing framework, one will decide in favor of the alternative model (reject the null hypothesis) only when the p-value is sufficiently small. Within the discrepancy function / model selection framework, the decision is based on which model is deemed to provide the more accurate depiction of the underlying data generating mechanism. In this paper, we establish a connection between the frameworks. We will show how the p-value can serve as an estimate of the probability on the null model under the discrepancy function framework. Furthermore, we will discuss the implications of imposing significance testing principles on a model selection problem.

Key Words: model selection, decision analysis, estimation, bootstrap

1. Introduction

Berger and Wolpert (1988) write that "Advancement of a subject usually proceeds by applying to complicated situations truths discovered in simple settings." Indeed, that is the hope for this paper. To motivate our ideas, we will focus on a simple problem. Suppose $Y_1, \ldots, Y_n, Y$ are independent and identically distributed with mean μ (unknown) and variance σ^2. The problem is to predict y after observing $y_1, \ldots, y_n$. Predicting the future based on observing the past is the essence of statistics. Our problem, although simple, is not trivial, and will provide a useful template to more complicated problems in statistical inference and model selection.

Let $\widehat{y}$ denote the predicted value of y. We will require a judgement on the accuracy of the prediction. Define

$$\begin{aligned} \Delta\left(\widehat{y}\right) &= E_Y\left(Y-\widehat{y}\right)^2 \\ &= \sigma^2 + \left(\widehat{y}-\mu\right)^2 \end{aligned}$$

as the mean squared error of prediction. Aside from the distribution variance, prediction accuracy under the mean squared error criterion depends only on the accuracy of $\widehat{y}$ as an estimate of the distribution mean μ. So for this problem, a prediction $\widehat{y}$ is synonymous with an estimate $\widehat{\mu}$. The choice of an estimator will depend on the selection of a model. The following models are under consideration:

$$\begin{aligned} M_o : Y_1, \ldots Y_n, Y &\sim N\left(\mu_o, \sigma^2\right) \\ M : Y_1, \ldots Y_n, Y &\sim N\left(\mu, \sigma^2\right). \end{aligned}$$

The null mean μ_o is a prespecified value. Under model M_o, we put forth the estimate $\widehat{\mu} = \mu_o$ regardless of the data observed. Under model M, we use the data information in creating the estimate $\widehat{\mu} = \overline{y}$. The decision between models is to be based on which model puts forth the more accurate estimator.

*Southern Illinois University Edwardsville, Department of Mathematics and Statistics, aneath@siue.edu

†The University of Iowa, Department of Biostatistics

‡The University of Iowa, Department of Biostatistics

The problem as stated is reminiscent of a hypothesis testing problem. Data $y_1, \ldots, y_n$ is to be used in testing the null hypothesis $H_o : \mu = \mu_o$ against a general alternative $H_a : \mu \neq \mu_o$. A decision in a null hypothesis significance testing problem may proceed through the use of a p-value. Define

$$\begin{aligned} p &= 2\left(1 - \Phi\left(|z|\right)\right) \\ &= \Pr\left(\chi^2\left(1\right) > z^2\right) \end{aligned}$$

where $z = \sqrt{n}\left(\overline{y} - \mu_o\right)/\widehat{\sigma}$ is the standardized test statistic. The problem is also reminiscent of model selection within a discrepancy function framework. Write

$$d\left(\widehat{\mu}, \mu\right) = \left(\widehat{\mu} - \mu\right)^2.$$

A discrepancy function provides a measure of disparity between the true model and a fitted candidate model. The mean squared estimation error d may be treated as a discrepancy function, where the fit of a model is judged by a comparison between the estimated mean and the true mean. The discrepancy d is the focus of inference in a model selection problem. However, d is a random variable since the fitted value $\widehat{\mu}$ is a function of the sample $Y_1, \ldots, Y_n$. We instead should say that the *distribution* on the discrepancy d is the quantity of interest.

In the next section, we establish a connection between the significance testing approach and the discrepancy function approach to model selection. We begin with the simple problem described here, then show how the development holds in a much broader setting. In Section 3, we consider two examples. The paper closes with some concluding remarks. The primary contribution of the paper is the presentation of a useful interpretation of a p-value. The connection works in both directions. We will also discuss the implications of imposing significance testing principles on a model selection problem.

2. P-value as an estimated probability

Under the null model M_o, the discrepancy $d\left(\mu_o, \mu\right)$ does not involve the observed sample. Its distribution is simply a point mass at $\left(\mu_o - \mu\right)^2$. Under general model M, the distribution on the discrepancy $d\left(\overline{Y}, \mu\right) = \left(\overline{Y} - \mu\right)^2$ is induced from the distribution on the sample mean. Model selection criteria are often developed by focusing on the expected value of an overall discrepancy. See McQuarrie, Tsai (1998) or Burnham, Anderson (2002) for an overview. Instead of summarizing the distributions via an expectation, we will base our model evaluation on the probability

$$\Pr\left[d\left(\mu_o, \mu\right) < d\left(\overline{Y}, \mu\right)\right] = \Pr\left[\left(\mu_o - \mu\right)^2 < \left(\overline{Y} - \mu\right)^2\right]. \tag{1}$$

So, the preferred model is that which is most likely to provide the more accurate estimate of the true distribution mean. Think of the null model discrepancy $\left(\mu_o - \mu\right)^2$ as a bias due to model misspecification. Think of the general model discrepancy $\left(\overline{Y} - \mu\right)^2$ as an error due to parameter estimation. When model bias is negligible in comparison to estimation error, the null model will be preferred. This may be so without the null conforming precisely to the truth. A fundamental aspiration of the discrepancy function approach to model selection is a balance between goodness of fit and parsimony.

We will use the bootstrap to estimate the distributions on the respective discrepancies. Let $\left(y_1^{(b)}, \ldots, y_n^{(b)}\right)$ denote a sample of size n from the empirical distribution. The sample

mean $\overline{y}$ serves as the empirical distribution mean. Let $\overline{y}^{(b)}$ denote the mean of the bootstrap sample. A bootstrap realization from the distribution on $d(\mu_o, \mu)$ is written as

$$\begin{aligned} \widehat{d}(\mu_o, \mu) &= d(\mu_o, \overline{y}) \\ &= (\mu_o - \overline{y})^2 . \end{aligned}$$

Since the distribution on $d(\mu_o, \mu)$ consists of a point mass only, so does its estimate. Denote a bootstrap realization from the distribution on $d\left(\overline{Y}, \mu\right)$ as

$$\begin{aligned} \widehat{d}(\overline{y}, \mu) &= d\left(\overline{y}^{(b)}, \overline{y}\right) \\ &= \left(\overline{y}^{(b)} - \overline{y}\right)^2 . \end{aligned}$$

Repeat for $b = 1, \ldots, B$ to create the estimated distribution on the discrepancy for the alternative model. Let $\Pr^*$ denote probability with respect to this estimated distribution. The bootstrap sampling scheme leads to an estimate of the probability in (1) as

$$\begin{aligned} \widehat{\Pr}\left[d(\mu_o, \mu) < d\left(\overline{Y}, \mu\right)\right] &= \Pr^*\left[\widehat{d}(\mu_o, \mu) < \widehat{d}\left(\overline{Y}, \mu\right)\right] \\ &= \Pr^*\left[(\mu_o - \overline{y})^2 < \left(\overline{Y}^{(b)} - \overline{y}\right)^2\right] \\ &\approx \frac{1}{B}\sum_{b=1}^{B} 1\left\{(\mu_o - \overline{y})^2 < \left(\overline{y}^{(b)} - \overline{y}\right)^2\right\} . \end{aligned} \tag{2}$$

We see in expression (2) the features which define the problem of deciding between two models. Support for the alternative model is strongest when the distance between the null mean and sample mean is large compared to the sampling variability. These same features appear when we take a significance testing approach to model selection. We are now in a position to present an argument connecting discrepancy function estimation and significance testing.

If the model assumptions on the true distribution are nearly correct, then the sampling distribution on the sample mean is approximately normal,

$$\overline{Y} \approx N\left(\mu, \frac{\sigma^2}{n}\right) .$$

Bootstrap resampling is ideal in the case when the empirical distribution captures the features of the true distribution. If the true distribution is approximated by the empirical distribution, then the bootstrap distribution on the sample mean is also approximately normal,

$$\overline{Y}^{(b)} \approx N\left(\overline{y}, \frac{\sigma^2}{n}\right) .$$

The bootstrap distribution on the overall discrepancy is induced to become

$$\left(\overline{Y}^{(b)} - \overline{y}\right)^2 \approx \frac{\sigma^2}{n}\chi^2(1) .$$

That is, the estimated distribution on the sampling error under the general model is a scaled chi-square distribution. We can take another look at the estimated probability in expression (2). Write

$$\begin{aligned} \widehat{\Pr}\left[d(\mu_o, \mu) < d\left(\overline{Y}, \mu\right)\right] &= \Pr^*\left[(\mu_o - \overline{y})^2 < \left(\overline{Y}^{(b)} - \overline{y}\right)^2\right] \\ &\approx \Pr\left[\frac{\sigma^2}{n}\chi^2(1) > (\mu_o - \overline{y})^2\right] \\ &= \Pr\left[\chi^2(1) > z^2\right] . \end{aligned}$$

Thus, we have the approximation

$$\widehat{\Pr}\left[d\left(\mu_o, \mu\right) < d\left(\overline{Y}, \mu\right)\right] \approx p. \tag{3}$$

Expression (3) establishes the p-value as an estimate of the probability that the null model provides a more accurate estimator than the general model.

Arguments and criticisms against the p-value are numerous, yet the use of the p-value in statistical inference has not slowed appreciably. Ideally, a statistical testing problem would be summarized through a measure of belief and/or a measure of evidence. Typically, a p-value does not provide such an interpretation. However, there are situations in which the p-value does provide a more desirable interpretation. Pawitan (2001) considers likelihood inference in the case when asymptotic normality holds and shows that the p-value provides information equivalent to a likelihood based measure of evidence. Goodman (2001) writes on interpreting a p-value by transforming p into a minimum Bayes factor. Casella, Berger (1987) show that the p-value matches a measure of belief in a Bayesian sense for a one-sided testing problem with a symmetric, noninformative prior.

By appealing to a discrepancy based model selection view of hypothesis testing, our result allows for an intriguing view of the p-value. We presented a simple case to aid in understanding, but the proof leading to (3) only requires a pivotal quantity, a normal approximation, and a discrepancy defined as a squared error of estimation. To see how the development holds in a broader setting, consider regression coefficient testing within a generalized linear model. Define competing models

$$\begin{aligned} M_o &: \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} \\ M &: \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \beta_k x_k \end{aligned}$$

where η is a function of the mean response. Deciding between models M_o and M is analogous to testing the null hypothesis $H_o : \beta_k = 0$. A significance testing approach is based on the p-value

$$p = \Pr\left[\chi^2\left(1\right) > w\right]$$

where $w = \left(\widehat{\beta}_k / SE_k\right)^2$ is a Wald statistic. We will compare this to a model selection approach based on the discrepancy

$$d\left(\widehat{\beta}_k, \beta_k\right) = \left(\widehat{\beta}_k - \beta_k\right)^2.$$

The quantity of interest in our model selection problem is the probability

$$\Pr\left[d\left(0, \beta_k\right) < d\left(\widehat{\beta}_k, \beta_k\right)\right] = \Pr\left[\left(0 - \beta_k\right)^2 < \left(\widehat{\beta}_k - \beta_k\right)^2\right]. \tag{4}$$

Model selection is based on the probability that the null model provides a more accurate estimate of the regression coefficient β_k than the alternate model. Again, it is not a requirement for the null model to be true for the null model to be better. It may be that setting the coefficient estimate to zero is more accurate than an estimate using data, because the general model estimate introduces additional sampling variability. An estimate of the probability in (4) computed via bootstrap resampling becomes

$$\widehat{\Pr}\left[d\left(0, \beta_k\right) < d\left(\widehat{\beta}_k, \beta_k\right)\right] = \frac{1}{B}\sum_{b=1}^{B} 1\left\{\left(0 - \widehat{\beta}_k\right)^2 < \left(\widehat{\beta}_k^{(b)} - \widehat{\beta}_k\right)^2\right\}$$

An argument completely analogous to the one made in the first problem establishes

$$\widehat{\Pr}\left[d\left(0,\beta_k\right) < d\left(\widehat{\beta}_k,\beta_k\right)\right] \approx p. \tag{5}$$

The derivation leading to (5) establishes the p-value as an estimate of the probability on the null model under a discrepancy function framework.

3. Examples and Discussion

3.1 Point spread data example

Gelman et al. (2003) present data on the scores of NFL football games. A point spread is designated for betting purposes prior to each game. The point spread represents the number of points added to the underdog (i.e., the team perceived to be weaker) score so that the dispersion of bets on each team is roughly the same. The bookmaker, taking a percentage off each bet, is guaranteed to make money provided that the point spread is a fair reflection of betting preference. We wish to determine whether or not there is some additional information beyond what is represented in the point spread. Define the actual point differential as the favored team's score minus the underdog score. The actual point differential will be negative if the weaker team surprises by winning the game. Define random variables

$$Y_1, \ldots Y_n \sim \left(\mu, \sigma^2\right)$$

where Y_i is the actual point differential minus the point spread. The point spread represents a single number summary for the entire pool of bettors. If bettors are exhibiting rational behavior, their actions should reflect all relevant information available prior to the game. Define the null model as $M_o : \mu = 0$. Call this the "wisdom of the crowd" model in that the information used by the pool of bettors is not exhibiting a systematic bias.

We are in the setting of the simple problem described in Section 1. The data consists of $n = 672$ games, resulting in sample mean $\overline{y} = 0.07$ and sample standard deviation $s = 13.86$. The p-value for testing $H_o : \mu = 0$ computes to be

$$\begin{aligned} p &= \Pr\left[\chi^2\left(1\right) > 672\left(\frac{0.07-0}{13.86}\right)^2\right] \\ &= .896. \end{aligned}$$

Null hypothesis significance testing, and by extension the p-value, is criticized for the illogical premise of testing the correctness of a theory when no theory is exactly correct. The discrepancy function approach fills the gaps in logic. Recall that the search is for the model providing the most accurate prediction / estimation. The null model may be best in this sense without matching the true model. Using the bootstrap algorithm for estimating the probability on the null, we get

$$\begin{aligned} \widehat{\Pr}\left[d\left(0,\mu\right) < d\left(\overline{Y},\mu\right)\right] &= \frac{1}{B}\sum_{b=1}^{B} 1\left\{(0-0.07)^2 < \left(\overline{y}^{(b)} - 0.07\right)^2\right\} \\ &= .892. \end{aligned}$$

The probability estimate is well approximated by the p-value. The discrepancy approach then allows for a nice interpretation of the p-value. Clearly the p-value is not the probability that the null model is true. But the p-value is, in some sense, the estimated probability that the null model is better than the general alternative. In the example, we do not believe the "wisdom of the crowd model" is precisely true. However, the large p-value can be

interpreted as providing strong support that this model is the best of those models available. The result here is consistent with the broader observation that betting markets regularly account for all relevant information. In fact, the point spread is of interest to even those fans not wagering on the games, as it provides one of the best measures on the relative strength of the competing teams.

3.2 Rhabdo data example

Exertional rhabdomyolysis (Rhabdo) occurs when strenuous exercise causes excessive skeletal muscle cell breakdown. Creatine kinase (CK) is a biomarker used for diagnosing Rhabdo. Smoot et al. (2014) provide data for studying the CK levels in Division I football players at the University of Iowa. The covariates include position, height, weight, age, and race. Our example will focus on the association between a pre-camp measurement of CK level and a measurement of CK level after 1 week of camp. CK levels are log transformed in our model since the distributions on the pre and post camp measurements are right skewed. The null model assumes no association between these two measurements. Under this model, a monitoring system for elevated CK levels would not need player specific baseline measurements. We compute the p-value for testing the corresponding null hypothesis $H_o : \beta_k = 0$ as

$$\begin{aligned} p &= \Pr\left[\chi^2(1) > \left(\frac{\widehat{\beta}_k - 0}{SE_k}\right)^2\right] \\ &= .0090. \end{aligned}$$

We compute a bootstrap estimate of the probability on the null model under a squared error discrepancy as

$$\begin{aligned} \widehat{\Pr}\left[d\left(0, \beta_k\right) < d\left(\widehat{\beta}_k, \beta_k\right)\right] &= \frac{1}{B}\sum_{b=1}^{B} 1\left\{\left(0 - \widehat{\beta}_k\right)^2 < \left(\widehat{\beta}_k^{(b)} - \widehat{\beta}_k\right)^2\right\} \\ &= .0092. \end{aligned}$$

Again we take note of the accuracy of the p-value as an approximation to a discrepancy function probability.

Under the significance testing philosophy, one does not decide in favor of the alternative (i.e., reject the null hypothesis) unless the p-value is sufficiently small. There is an appeal to this sort of thinking. Philosophy of science reflects a preference toward the null model. For instance, Occam's Razor and Popper's view on falsifiability are both founded on the principle of building from simple to more complex models. Good science calls for the acceptance of the simplest model which provides an accurate representation of the observed data. Because of the small p-value in the Rhabdo example, the null model is not considered to be acceptable. Therefore, a decision in favor of the more general alternative is justified.

Under our discrepancy function framework, one prefers the model most likely to put forth the more accurate estimate. Because this probability is estimated by the p-value, the rule is to decide in favor of the alternative when $p < 1/2$. This is quite a different viewpoint from the significance testing philosophy. But we can use the information from p to perform a model evaluation rather than simply a model selection. The computation of a small probability on the null model offers a separation between a case where we merely select the more general alternative and a case where the data is providing a clear distinction. A requirement of a small probability on the null before deciding in favor of the larger model is in line with the significance testing philosophy of favoring the null model unless a strong

indication to the contrary is provided by the data. Here, the evidence is substantial in favor of the model which includes an association between the baseline CK level and the week 1 CK level. Behind the principle requiring p to be sufficiently small is the premise that one is not willing to recognize the larger model until sufficiently convinced that the regression coefficient (effect size) can be distinguished from zero using the available data.

4. Concluding Remarks

The results in this paper are limited to the use of a squared error discrepancy and a candidate class with only two (nested) models. Such conditions are necessary to form a connection between discrepancy function estimation and the p-value. We must emphasize, however, that these conditions are not necessary for the use of probability measurements as an evaluation on a candidate class of models. Neath, Cavanaugh, and Riedle (2012) use the bootstrap resampling scheme described in this paper for model selection problems based on the Kullback-Leibler discrepancy with no restrictions on the candidate class.

In this paper, we have shown that the p-value may be interpreted as the probability on the null model providing a more accurate estimate than a general alternative. The interpretation holds for a selection between two models based on a discrepancy defined by the squared error of estimation. The p-value, although widely used, suffers from a lack of interpretability. The result provided in this paper allows one to better understand the information provided by a p-value summary. Furthermore, the use of probability in a discrepancy function framework allows one to measure the support for a model using a very familiar criterion. A small p-value, as would be required for accepting the larger model in a significance testing framework, indicates a clear decision under the model selection framework.

REFERENCES

Berger, J., and Wolpert, R. (1988), *The Likelihood Principle: A Review, Generalizations, and Statistical Implications*, Hayward, CA: IMS Lecture Notes

Casella, G. and Berger, R. (1987), "Reconciling Bayesian and frequentist evidence in the one-sided testing problem," *Journal of the American Statistical Association*, 82, 106-111.

Burnham, K., and Anderson, D. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, New York: Springer.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, London: Chapman and Hall.

Goodman, S. (2001) "Of p-values and Bayes: A modest proposal," *Epidemiology*, 12, 295-297.

McQuarrie, A. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*, River Edge, NJ: World Scientific.

Neath, A., Cavanaugh, J., and Riedle, B. (2012), "A bootstrap method for assessing uncertainty in Kullback-Leibler discrepancy model selection problems," *Mathematics in Engineering, Science, and Aerospace*, 3, 381-391.

Pawitan, Y. (2001), *In All Likelihood: Statistical Modeling and Inference Using Likelihood*, Oxford: University Press.

Smoot, M.K., Cavanaugh, J., Amendola, A., West, D., and Herwaldt, L. (2014), "Creatine kinase levels during preseason camp in National Collegiate Athletic Association Division I Football Athletes," *Clinical Journal of Sports Medicine*, 24, 438-440.