International Statistical Review (2005), **73**, 3, 351–363. Printed in Wales by Cambrian Printers
© International Statistical Institute

How to Choose a Working Model for Measuring the Statistical Evidence About a Regression Parameter

Jeffrey D. Blume

Center for Statistical Sciences, Brown University, Providence RI 02912, USA
E-mail: jblume@stat.brown.edu

Summary

Consider using a likelihood ratio to measure the strength of statistical evidence for one hypothesis over another. Recent work has shown that when the model is correctly specified, the likelihood ratio is seldom misleading. But when the model is not, misleading evidence may be observed quite frequently. Here we consider how to choose a working regression model so that the statistical evidence is correctly represented as often as it would be under the true model. We argue that the criteria for choosing a working model should be how often it correctly represents the statistical evidence about the object of interest (regression coefficient in the true model). We see that misleading evidence about the object of interest is more likely to be observed when the working model is chosen according to other criteria (e.g., parsimony or predictive accuracy).

Key words: The law of likelihood; Statistical evidence; Misleading evidence; Model selection.

1 Introduction

Every group of observations represents statistical evidence about the probability distribution that generated them. We learn about that distribution by properly examining and interpreting those data as statistical evidence. The mathematical representation of that evidence is the likelihood function (Birnbaum, 1962; Berger & Wolpert, 1988), and Law of Likelihood explains how the evidence should be measured and interpreted (Hacking, 1965; Royall, 1997; Blume, 2002). Implicit here is the specification of a probability model, so that the interpretation and measurement of the evidence is, in some sense, conditional on that model.

In this paper we describe what happens when the regression model we use in our statistical analysis, call it the 'working' model, is not the 'true' model that actually generated the data. We will judge the performance of a working model by how often it produces misleading and weak evidence, both in absolute magnitude and relative to how often such evidence is observed when the model is correctly specified. Some working models will, in general, do a better job of correctly representing the statistical evidence than others. Our objective is to identify those models and provide some guidelines on how to choose a working regression model for representing and interpreting the statistical evidence about a parameter of interest.

This is a model selection problem where the criteria for selection is evidential performance, rather than predictive ability or parsimony. The literature on "regression model selection" is vast, with many different criteria and selection strategies having been suggested (e.g., Burnham & Anderson, 1998; Harrell, 2001). We will not attempt to summarize that literature here, but in passing note that the most frequently employed criteria are measures of predictive ability (e.g., AIC or BIC) or model fit (e.g., loglikelihood or deviance statistics). Interestingly, there is no immediate reason why a working model selected according to these (non-evidential) criteria should also correctly represent the statistical evidence about the regression coefficients of interest. In fact, we will see that these models have the potential to frequently misrepresent the statistical evidence.

In what follows section 1.1 briefly reviews the necessary background on the Likelihood paradigm. Section 2 considers the case of a normal linear regression, where we develop some criteria for selecting a working model. Then, in section 3, we investigate possible connections with other model selection criteria, such as AIC. Extensions, such as relaxing the normality assumption or allowing missing covariates, are mentioned in sections 4 and 5. Lastly, a summary discussion is given in section 6.

1.1 Background

If $y_1, \ldots, y_n$ are realizations of random variables $Y_1, \ldots, Y_n$ iid $f(Y_i;\theta)$, then $L(\theta) \propto \prod f(y_i;\theta)$ is the likelihood function and the likelihood ratio, $L(\theta_1)/L(\theta_2)$ measures the strength of the evidence supporting the hypothesis that $H_1 : \theta = \theta_1$ over $H_2 : \theta = \theta_2$ (Hacking, 1965; see also Birnbaum, 1962). The probability of observing strong evidence supporting a false value of theta, say $\theta \neq \theta_0$, over the true value of theta, θ_0, by a factor of k or greater is given by $P(L(\theta)/L(\theta_0) \geq k)$, which we refer to as the probability of observing k-strength misleading evidence or the probability of misleading evidence for short. Likewise, the probability of weak evidence is given by $P(1/k < L(\theta)/L(\theta_0) < k)$ and the probability of strong evidence by $P(L(\theta_0)/L(\theta) \geq k)$. Here $k > 1$ is a constant where the values of $k = 8$ and 32 represent fairly strong and strong evidence (Royall, 1997). See Royall (1997) for an illuminating and comprehensive discussion of this approach and its advantages over standard frequentist and Bayesian methodologies. An article by Blume (2002) provides an introduction to the Likelihood paradigm and overview of the key concepts.

Royall & Tsou (2003) recognized that, in this context, the likelihood function has two key performance properties:

1. For any false value $\theta \neq \theta_0$ the evidence will eventually support θ_0 over θ by an arbitrarily large factor: $P(L(\theta_0)/L(\theta) \to \infty \text{ as } n \to \infty) = 1$;
2. In large samples the probability of misleading evidence, $P(L(\theta)/L(\theta_0) \geq k)$, as a function of θ, is described by the bump function, $\Phi[-\ln k/c - c/2]$ where k is any constant greater than one, $\Phi[\cdot]$ is the standard normal distribution function and c is proportional to the distance between θ and θ_0, (Pratt, 1977; Royall, 2000).

The first property is important because it states that the evidence will eventually favor the truth over any fixed false alternative. This also implies that the probabilities of observing misleading and weak evidence converge to zero as the sample size increases.

Unlike the first property, which applies to all models, the second only applies to models where the likelihood function is sufficiently smooth. The second property states that, in large samples, the maximum probability of observing misleading evidence, over the parameter space, is $\Phi[-\sqrt{2 \ln k}]$. Here the alternative approaches the true value at some rate, rather than remaining fixed. In fact, c is most easily interpreted as the distance between the alternative and true hypothesis in standard error units (Royall, 2000; Blume, 2002).

In addition, we know that for any fixed sample size the probability of misleading evidence is bounded above by $1/k$. This bound is often referred to as the 'universal bound' because it applies under any probability model (Barnard, 1947; Smith, 1953; Birnbaum, 1962). However, in large samples, the maximum value of the bump function is typically much lower than the universal bound (e.g., when $k = 8$ we have 0.021 for the maximum of the bump function versus 0.125 for the universal bound).

It follows from the second property that in large samples, the probability of strong evidence is $1 - \Phi[\ln k/c - c/2]$ and the probability of weak evidence is $\Phi[\ln k/c - c/2] - \Phi[-\ln k/c - c/2]$. The dotted lines in figure 1 represent these three probabilities when $k = 8$. The x-axis is the distance between the true parameter and the alternative in standard error units. The variables 'q' and 'c_K' will be defined and discussed in the next section.

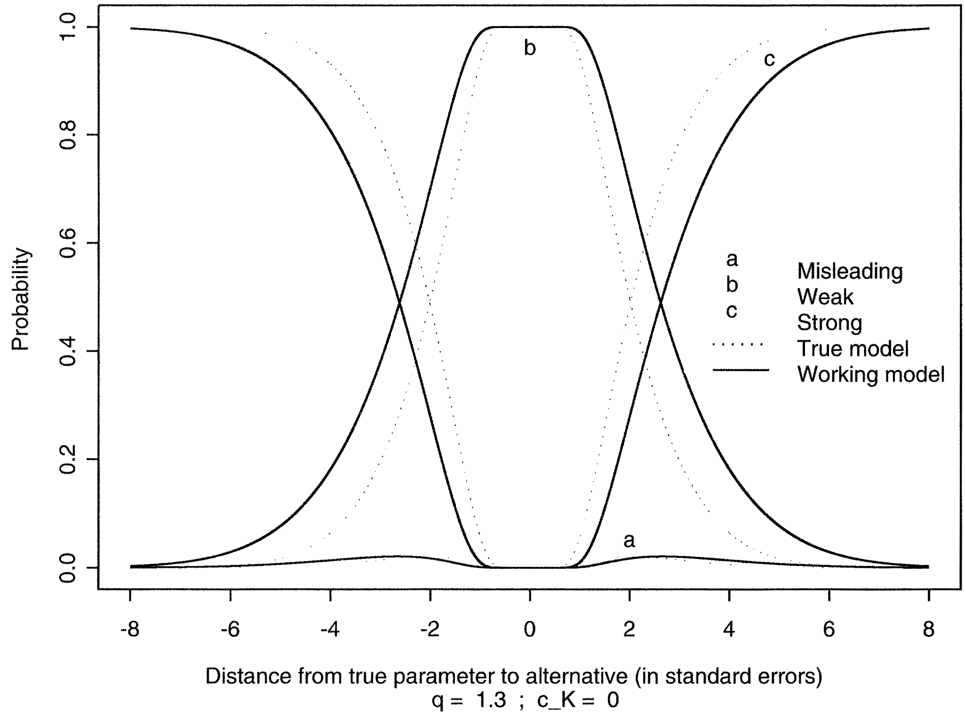

Figure 1. *Probabilities of misleading, weak and strong evidence when the working model is the large model and the true model is the small model.*

2 Regression Models

We are interested in how one variable tends to change when another one does. Let $Y = (Y_1, \cdots, Y_n)$ be a $n \times 1$ vector of independent responses, X be the $n \times 1$ vector for the covariate of interest, and Z be a $n \times r$ matrix of additional covariates. Throughout this section we will consider the following two models:

$$
\begin{array}{lrcll}
(M_s) & Y & = & X\beta_s + e & \text{where } e \sim N(0, \sigma^2 I_n) \\
(M_l) & Y & = & X\beta_l + Z\gamma + e & \text{where } e \sim N(0, \sigma^2 I_n)
\end{array}
$$

where β_s and β_l are scalar parameters, $\gamma \neq 0$ is a $r \times 1$ parameter vector, and σ^2 is an unknown constant. Without any loss of generality, we will assume that the responses y and covariates x and z_j ($j = 1, \ldots, r$) are centered (this centering simply forces the intercept to be zero). Their respective sum of squares will be represented by $S_{yy} = \sum(y_i - \overline{y})^2$, $S_{xx} = \sum(x_i - \overline{x})^2$ and $S_{jj} = \sum_i (z_{ji} - \overline{z_j})^2$.

We will refer to these two models as the small and large model respectively.

We have used subscripts on β to emphasize that these parameters are, in general, not equal. Following Royall & Tsou (2003), we differentiate between the parameter in the true model that we *wish* to learn about (the object of interest) and the parameter in the working model that we *will* learn about (the object of inference). Here the object of interest is the coefficient of X in the true regression model, while the object of inference is the coefficient of X in the working regression model. Ideally, the object of inference will represent the same effect that the object of interest does, but this may not be the case when the working model does not contain the true model.

Our goal is to determine how well particular working models characterize the statistical evidence about the object of interest. That is, how is the strength of evidence affected by our choice of the working model? We answer this question by calculating the probabilities of observing misleading and weak evidence when using these working models and evaluating how the probabilities change from the expressions given in the previous section.

One benefit of this simple distributional structure is that we can easily obtain a one dimensional marginal likelihood for each regression coefficient. However, when the errors are not normally distributed an analogous marginal likelihood may not be so readily attainable. One solution is to use a profile likelihood function and this is discussed later (section 4).

2.1 When the Large Model is our Working Model

Here we examine the effect of including additional covariates Z in our working regression model when these covariates are really unimportant (i.e., $\gamma = 0$ and M_s is the true model). Now β_s is the coefficient of interest and we learn about β_s by examining $L_l(\beta)$, the marginal likelihood function for β under the large model. We will use the working likelihood ratio $L_l(\beta_1)/L_l(\beta_2)$ to measure the strength of evidence supporting the hypothesis that the true coefficient is β_1 versus the hypothesis that it is β_2.

With the error variance unknown, it is sufficient to work with the following approximation to t-distribution:

$$L_l(\beta) \propto \exp\left\{-\frac{(\widehat{\beta}_l - \beta)^2}{2V_{l|s}}\right\} \tag{1}$$

where $\widehat{\beta}_l$ is the least squares/maximum likelihood estimator (MLE) from the working model, $\widehat{V}_l = \widehat{\mathrm{Var}}_l[\widehat{\beta}_l]$ is the estimated variance of $\widehat{\beta}_l$ under the working model and $V_{l|s} = E_s[\widehat{V}_l]$ is the expected value of that working variance under the true model. Here we have used the vertical bar in the subscript to emphasize that these calculations are conditional on the true model.

Some details are necessary before proceeding. $\widehat{\beta}_l$ is given by the first component of $(W^T W)^{-1} W^T Y$ where $W = [X \;\; Z]$. Routine calculations show that the estimate of its variance is $\widehat{V}_l = MSE_l[S_{xx}(1 - R_x^2)]^{-1}$ where MSE_l is the mean squared error from the large model and R_x^2 is the coefficient of multiple determination that results from regressing X on the columns of Z (Neter, Wasserman, Kutner, 1990 p. 409; Montgomery & Peck, 1992 p. 309; Seber, 1977).

The law of large numbers implies that $\widehat{V}_l \to E_s[\widehat{V}_l] = V_{l|s}$ and because MSE_l remains unbiased for σ^2 under the small model, i.e. $E_s[MSE_l] = \sigma^2$, we have that $V_{l|s} = \sigma^2[S_{xx}(1 - R_x^2)]^{-1}$. The term $(1 - R_x^2)^{-1}$ is known as the variance inflation factor. If Z is a vector containing only one additional covariate, say z, then $R_x^2 = r_{xz}^2$, the square of the correlation between x and z. Thus, including covariates in Z that are correlated with the regressor of interest X increases the variance.

What does this tell us about the probability of observing misleading or weak evidence? Because the small model is true, $E_s[\widehat{\beta}_l] = \beta_s$ and the law of large numbers implies that property one is satisfied. Namely, for any false value $\beta \neq \beta_s$ the evidence will eventually support β_s over β by an arbitrarily large factor,

$$P_s(L_l(\beta_s)/L_l(\beta) \to \infty \text{ as } n \to \infty) = 1.$$

Be careful to take note of the subscripts here; both the probability space and the object of interest are indexed by the true model while the likelihood function is specified by the working model.

In addition to the first property, the second property is also satisfied: the probability of observing misleading evidence under this working model follows the bump function. In large samples, this probability is

$$P_s\left(\frac{L_l(\beta)}{L_l(\beta_s)} \geq k\right) = \Phi\left[-\frac{q\ln k}{|c|} - \frac{|c|}{2q}\right] \tag{2}$$

where c is proportional to the distance between β and β_s, and $q = \sqrt{V_{l|s}/V_{s|s}}$ denotes the ratio of variances where $V_{s|s} = E_s[\widehat{\text{Var}}_s[\widehat{\beta}_s]] = \sigma^2[S_{xx}]^{-1}$ is the variance of the small model least squares estimator $\widehat{\beta}_s$ (see the appendix).

The factor $q = (1 - R_x^2)^{-1/2} \geq 1$ does not change the general shape of the bump function: it merely spreads the curve out along the x-axis. The maximum probability of misleading evidence, over the parameter space, is still given by $\Phi[-\sqrt{2\ln k}]$. If we were to compare the working likelihood function with the true one, we would see that the working likelihood function is flatter and that the evidence is therefore weakened. However if $R_x^2 = 0$, meaning that none of the additional covariates are linearly related to the regressor of interest X, then both likelihood functions will have the same curvature.

It follows that, in large samples, the probability of strong evidence is $1 - \Phi[q\ln k/c - c/2q]$ and the probability of weak evidence is $\Phi[q\ln k/c - c/2q] - \Phi[-q\ln k/c - c/2q]$. This is illustrated by the solid lines in figure 1 when $k = 8$ and $q = 1.3$ (i.e., $R_x^2 = 0.4$). Note that if $R_x^2 = 0$ the solid lines would lie over the dotted lines. Note the loss in efficiency: the probability of strong evidence drops at every alternative, while the probability of weak evidence increases at every alternative.

2.2 *When the Small Model is our Working Model*

Now we examine the effect of omitting the additional covariates Z from our working regression model when these covariates are really important (i.e., $\gamma \neq 0$ and M_l is the true model). Here the coefficient of interest is β_l, which represents the expected change in response, Y, per unit change in the regressor x, *all other things being equal* (i.e., z being held fixed).

Analogous arguments to the previous section apply here, so that a good approximation to the working likelihood function is

$$L_s(\beta) \propto \exp\left\{-\frac{(\widehat{\beta}_s - \beta)^2}{2V_{s|l}}\right\} \tag{3}$$

where $\widehat{\beta}_s$ is the least squares/maximum likelihood estimator (MLE) from the working model, $\widehat{V}_s = \widehat{\text{Var}}_s[\widehat{\beta}_s]$ is the estimated variance of $\widehat{\beta}_s$ under the working model and $V_{s|l} = E_l[\widehat{V}_s]$ is the expected value of that working variance under the true model.

The estimate of β_l from the working model is $\widehat{\beta}_s = (X^TX)^{-1}X^TY$ and its estimated variance is $\widehat{V}_s = MSE_s[S_{xx}]^{-1}$ where MSE_s is the mean squared error from the small model. But here the working mean squared error overestimates σ^2. That is, $E_l[MSE_s] = \sigma^2 + B$ where $B = \gamma^T Z^T(I_n - P_s)Z\gamma/(n - r - 1)$ is a bias term and $P_s = X(X^TX)^{-1}X^T$ is the projection matrix from the small model. The bias term reduces to

$$B = \sum_{j=1}^{r} \gamma_j\sigma_{jj} \sum_{i=1}^{r} \gamma_i\sigma_{ii}(r_{ij} - r_{ix}r_{jx}) \tag{4}$$

where r_{ix} is the correlation between X and the i-th column/covariate in Z, γ_i is the i-th component of the parameter vector γ and $\sigma_{jj} = \sqrt{S_{jj}/(n - r - 1)} = O_p(1)$ is a constant.

Now, by the law of large numbers, $\widehat{V}_s \rightarrow V_{s|l} = (\sigma^2 + B)[S_{xx}]^{-1}$. Because $\gamma \neq 0$ when large model is true, the bias term is always greater than zero. This is because when $i = j$ the contribution to the bias in the MSE is $\gamma_i^2\sigma_{ii}^2(1 - r_{ix}^2)$, which is zero only when the regression cannot provide unique estimates due to perfect collinearity of some covariates. Then $B = \gamma^2\sigma_{zz}^2(1 - r_{xz}^2)$, which is zero only when X and Z are perfectly correlated.

The evidential consequences of using the small working model are quite drastic. To begin with, the first property no longer applies, so we are no longer assured that the evidence will eventually favor the object of interest over any false value. This should not be surprising given that the object of inference, $\widehat{\beta_s}$, is generally biased and not consistent for the object of interest, β_l. Appealing to the law of large numbers, we have that $\widehat{\beta_s} \rightarrow E_l[\widehat{\beta_s}] = \beta_l + K\gamma$ where the $1 \times r$ vector $K = (X^TX)^{-1}X^TZ = [r_{x1}\sqrt{S_{11}/S_{xx}} \cdots r_{xr}\sqrt{S_{rr}/S_{xx}}]$ is a vector of scaled correlations. The bias in estimating the object of interest disappears if the regressor of interest X is uncorrelated with each additional covariate in Z.

In fact, when $K \neq 0$, we are *assured* of observing misleading evidence about the object of interest, β_l. Specifically we have that for any value $\beta \neq (\beta_l + K\gamma)$ the evidence will eventually support $\beta_l + K\gamma$ over β by an arbitrarily large factor. Setting β equal β_l demonstrates that we cannot (reliably) learn about the object of interest from the small working model in this case.

This is a situation where the object of inference $(\beta_l + K\gamma)$ does not, in general, equal the object of interest (β_l). Hence when $K \neq 0$, the small working model will produce strong misleading evidence about β_l with probability one. But when the additional covariates are uncorrelated with the regressor X (i.e., $r_{ix} = 0$ for all i and therefore $K = 0$), the first property holds and we can reliably learn about the object of interest (although with some loss of efficiency). *Thus it is critical to include important covariates (i.e., those with a truly nonzero regression coefficient) in the working model when they are correlated with the regressor X in order to maintain validity.*

What about the second property? Does the probability of observing misleading evidence under this working model follow the bump function in large samples? Not quite. In large samples, the probability of observing misleading evidence is given by

$$P_l\left(\frac{L_s(\beta)}{L_s(\beta_l)} \geq k\right) = \Phi\left[-\frac{q\ln k}{|c|} - \frac{|c|}{2q} + \text{sign}(c)\frac{c_K}{q}\right] \tag{5}$$

where the scaled bias, $c_K \propto K\gamma$, is in standard error units and equals zero when $K = 0$, sign$(c) = 1$ if $c > 0$ and -1 if $c < 0$ and $q = \sqrt{V_{s|l}/V_{l|l}}$ denotes the ratio of variances (see the appendix for proof). Here $q = \sqrt{(1 - R_x^2)(1 + \frac{B}{\sigma^2})} \geq 0$.

When the first property does not hold, the maximum probability of misleading evidence over the parameter space, $\Phi[-\sqrt{2\ln k} + |c_K|/q]$, is unconstrained and depends on the magnitude of bias, $K\gamma$ and the sample size. In addition, it is also possible that we gain efficiency over the true model (i.e., $q < 1$). This happens when $K\gamma \neq 0$ and $R_x^2 > B/(B + \sigma^2)$; that is when the bias, B, is small relative to the overall variance σ^2. This is an example of the bias/variance tradeoff.

The previous maximum of the bump function, $\Phi[-\sqrt{2\ln k}]$, holds only when $K = 0$. And if $K = 0$ then $R_x^2 = 0$, and $q \geq 1$. Therefore, when property one is satisfied, we cannot learn about the object of interest more efficiently by using a misspecified working model.

It follows that, in large samples, the probability of strong evidence is $1 - \Phi[q\ln k/c - c/2q + \text{sign}(c)c_K/q]$ and that the probability of weak evidence is $\Phi[q\ln k/c - c/2q + \text{sign}(c)c_K/q] - \Phi[-q\ln k/c - c/2q + \text{sign}(c)c_K/q]$. This is illustrated in figure 2 when $k = 8$, $q = 1.3$ (i.e., $R_x^2 = 0.4$) and $c_K = -1.3$. Notice the distortion of the solid curves relative to the dotted curves; this is mainly due to the fact that the first property is not satisfied here. Also, the probability of observing misleading evidence does not appear to approach one because here we are allowing the alternative to approach the true value as the sample size increases.

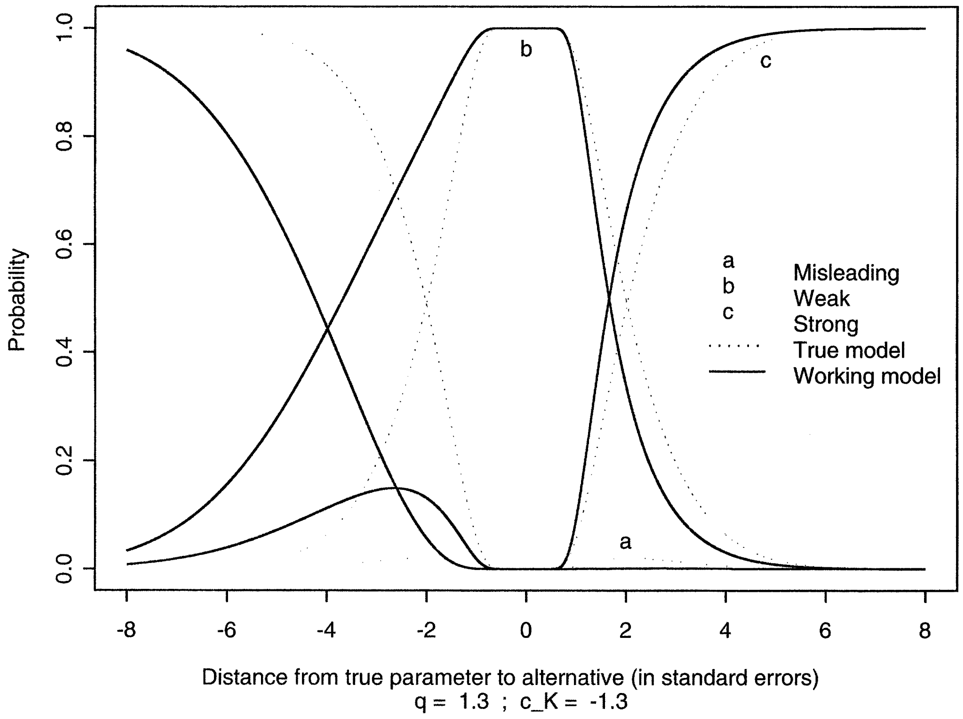

Figure 2. *Probabilities of misleading, weak and strong evidence when the working model is the small model and the true model is the large model.*

2.3 The Effect of Including Important, Uncorrelated Regressors

As in the previous section, the large model is the true model and we use the small model as the working model. We just saw that if the additional covariates are uncorrelated with the regressor X (i.e., $K = 0$) both of the desired properties on the probability of misleading evidence are satisfied by the small working model. Why then, should we bother to include these covariates in the regression equation?

The answer is "for efficiency", or from a likelihood perspective, "to uniformly increase the probability of strong evidence". That is, we reduce q. When $K = 0$ the probability of strong evidence is

$$P_l\left(\frac{L_s(\beta_l)}{L_s(\beta)} \geq k\right) = 1 - \Phi\left[\frac{q \ln k}{c} - \frac{c}{2q}\right] \tag{6}$$

where $q = \sqrt{1 + B/\sigma^2}$. Now B/σ^2 is the amount by which the MSE_s overestimates σ^2 and this bias term is reduced each time an important covariate is added to the small model. Therefore we include important covariates that are uncorrelated with the regressor of interest (x) because they increase the probability of strong evidence (i.e., they move q towards one). Remember that we leave important, correlated covariates in the working model to maintain validity, in the sense that property one is satisfied.

2.4 The Effect of Excluding Unimportant, Correlated Regressors

Consider the reverse situation from the preceding section: the small model is the true model and we use the large model as the working model. In this situation we should remove unimportant covariates that are correlated with the regressor of interest because doing so uniformly increases the probability of observing strong evidence. That is, we gain efficiency and reduce q. The probability of strong evidence is $1 - \Phi\left[\frac{q \ln k}{c} - \frac{c}{2q}\right]$, where $q = (1 - R_x^2)^{-1/2}$. Now R_x^2 will be reduced by removing covariates from the large model, which in turn reduces q. Therefore by removing unimportant, correlated covariates from the working model we increase the probability of strong evidence (i.e., we move q towards one).

3 Model Selection by Predictive Accuracy

Suppose we select for the working model the one which appears to be the most predictively accurate. Is the maximum probability of observing misleading evidence bounded under this model? Does the bump function represent the probability of observing misleading evidence under this model in large samples? The answer to both questions is that sometimes it does and sometimes it does not.

Consider selecting the working model with the smallest Akaike Information Criterion (AIC) (Akaike, 1973). There are many different selection criteria like this, but almost all of them amount to minimizing the residual sum of squares plus a penalty for overfitting (Zhang, 1992). For our purposes it suffices to consider only AIC here. Specifically, we want to know whether the working model with the smallest AIC will represent the evidence about β correctly.

The AIC is $-2\log(L(\widehat{\theta}|Y)) + 2p$ where $\widehat{\theta}$ is the vector of maximum likelihood parameter estimates and p is the total number of estimated parameters, including the variance. For the large model (M_l), the AIC reduces to $n + n\log(2\pi\widehat{\sigma_l}^2) + 2(r+2)$ where $\widehat{\sigma_l}^2 = SSE_l/n$ and $SSE_l = \sum(y_i - \widehat{y_l})^2$ is the residual sum of squares from the large model and $\widehat{y_l}$ are the fitted values from the large model. If SSE_s is the residual sum of squares from the small model then $SSE_l/SSE_s = 1 - R^2_{yz.x}$ where $R^2_{yz.x}$ is the correlation between Y and Z after adjusting both variables for X. $R^2_{yz.x}$ is also known as the partial correlation between Y and Z or the coefficient of partial determination.

Selecting for the working model that which has the smaller AIC leads to selecting the small model whenever the coefficient of partial determination, $R^2_{yz.x}$, is itself sufficiently small. Specifically, this happens whenever $AIC_l - AIC_s = n\log(1 - R^2_{yz.x}) + 2r > 0$ or whenever

$$R^2_{yz.x} < 1 - \exp\{-2r/n\}. \tag{7}$$

Notice that $R^2_{yz.x}$ is a random variable, so that the model selection process is also random. And when the correct model is not chosen, we have the consequences outlined in section 2. It can be shown that the probability of AIC choosing the correct model is at least 0.712 (Zhang, 1992 p.736; Woodroofe, 1982 p.1184). For example, when Z is a vector containing only one additional covariate ($r = 1$) and the small model is the true model, the probability that AIC selects the small model is approximately 0.84. As the number of (unimportant) additional covariates (r) increases, this probability tends to 0.712.

To illustrate that this type of selection process can be problematic from an evidential point of view, consider the case when $r = 1$. Then we have that $\widehat{\gamma} \propto R_{yz.x}$ and AIC excludes the covariate if its estimated coefficient is small enough. Here the coefficient of partial determination is

$$R^2_{yz.x} = \frac{(r_{yz} - r_{xz}r_{yx})^2}{(1 - r^2_{xz})(1 - r^2_{yx})} \tag{8}$$

and $R^2_{yz.x} = 0$ implies that $r_{xz} = r_{yz}/r_{yx}$. Thus Z adds nothing to the model because $|r_{yz}| < |r_{yx}|$. However the absolute magnitude of r_{xz} may still be large and the true coefficient of Z may be nonzero, resulting in non-ignorable increases in the probability of observing misleading evidence. And as $r_{xz} \to 1$, we have that $R^2_{yz.x} \to 0$ because $r_{yz} \to r_{yx}$, showing that AIC tends to exclude highly correlated covariates.

Model selection by AIC tends to exclude covariates that are highly correlated because they do not significantly increase the working model's ability to predict the response. However if those covariates are really important (i.e., $\gamma \neq 0$), then this will render the representation of the statistical evidence strongly and frequently misleading.

4 Extensions

Our treatment of this problem has assumed that the marginal distribution of the regression parameter estimate was known. When the model is correctly specified, we saw that the two key performance properties outlined in section 1.1 hold. But in many multiparameter models, such as $X_1, X_2, \ldots, X_n$ i.i.d. $f(X_i; \beta, \gamma)$, such conditional or marginal likelihoods are not known and other approaches must be identified.

One such approach is to use a profile likelihood function. The profile likelihood function maximizes the joint likelihood with respect to the nuisance parameter at each value of the parameter of interest. For fixed β, the profile likelihood function is defined as $\max_\gamma L_n(\beta, \gamma) = L_n(\beta, \widehat{\gamma}(\beta)) = L_p(\beta)$. We use the profile likelihood function for β as if it were a true likelihood function.

Profile likelihoods are an attractive option because if both β and γ are fixed dimensional parameters and f is a smooth function, then the profile likelihood will behave like a true likelihood in large samples. That is, for a profile likelihood ratio, the limiting probability of observing misleading evidence is given by the Bump function (Royall, 2000). We need only be sure that the object of interest is the object of inference and this happens, in general, when working model contains the true model.

Moreover, if the distribution $f(\cdot)$ is not correctly specified, there are ways to adjust the profile likelihood function so that it obeys the second property. Royall & Tsou (2003) show how this can be done in cases where the first property holds (i.e., the object of interest remains the object of inference). They briefly consider the case of profiling out the intercept term in a simple linear regression, but their approach is readily generalizable.

For example, our large working model implies that $Y \sim N(X\beta_l + Z\gamma, \sigma^2 I_n)$ and the profile likelihood for β_l is given by

$$\begin{aligned} L_p(\beta) &\propto (\widehat{\sigma}^2(\beta))^{-n/2} \times \exp\left[-\frac{(Y - X\beta - Z\widehat{\gamma}(\beta))^T(Y - X\beta - Z\widehat{\gamma}(\beta))}{2\widehat{\sigma}^2(\beta)}\right] \\ &\propto \left[\frac{(Y - X\beta - Z\widehat{\gamma}(\beta))^T(Y - X\beta - Z\widehat{\gamma}(\beta))}{n}\right]^{-n/2} \end{aligned}$$

where $\widehat{\gamma}(\beta) = (Z^T Z)^{-1} Z^T (Y - X\beta)$ and $\widehat{\sigma}^2(\beta) = (Y - X\beta - Z\widehat{\gamma}(\beta))^T(Y - X\beta - Z\widehat{\gamma}(\beta))/n$. Notice that the residual vector $R = Y - X\widehat{\beta} - Z\widehat{\gamma}(\widehat{\beta})$ can be re-expressed as $R = (Y - Z(Z^T Z)^{-1} Z^T Y) - (X - Z(Z^T Z)^{-1} Z^T X)\widehat{\beta}$, to show that the profile likelihood is just the likelihood obtained from regressing the 'corrected' covariates of interest $X^* = X - Z(Z^T Z)^{-1} Z^T X$ on the 'corrected' response $Y^* = Y - Z(Z^T Z)^{-1} Z^T Y$.

In this case the 'robust' profile likelihood is given by $[L_p(\beta)]^E$ where

$$E = \frac{X^{*T} X^* R^T R/n}{X^{*T} \mathrm{diag}\{R^T R\} X^*} = \frac{(X^{*T} X^*)^{-1} R^T R/n}{(X^{*T} X^*)^{-1} X^{*T} \mathrm{diag}\{R^T R\} X^* (X^{*T} X^*)^{-1}} .$$

So the exponential adjustment factor E is simply the ratio of the model based variance estimate for β to the robust variance estimate (sandwich estimator), but using the 'corrected' data. Note that $X^{*T} X^*$ is 1×1.

Profile likelihoods adjusted in this fashion will, in large samples, satisfy the second key performance property (even when the underlying distribution is not normal) as long as the first property is also satisfied (Royall & Tsou, 2003). A sufficient condition for the first property to hold is that the mean structure is correctly specified, but as we saw earlier the mean structure may omit Z as long as it is uncorrelated with X, the regressor of interest. It follows that, in large samples, the profile likelihood's behavior with respect to the omission of covariates will be similar to that described in the previous sections.

5 What if Important, Correlated Covariates are Missing?

It is sometimes the case that an important covariate is missing. If that missing covariate is uncorrelated with the regressor of interest, then at worst we will lose some efficiency. But if the missing covariate is correlated with the regressor of interest, then the evidence interpreted under our working model will be strongly misleading. Unfortunately when this happens there is little we can do to enable us to learn about the object of interest.

The situation is this: suppose that our large model $E(Y_i) = \alpha + \beta x_i + \gamma z_i$ for some $\gamma \neq 0$ is the true model, but covariate Z is missing. Here β remains our object of interest and the true model can be re-expressed as $E(Y_i) = \alpha_i + \beta x_i$. The difference between this model and our small working model is now clear: there is no longer a common intercept. This is problematic because the number of parameters is now always greater than the number of observations. A profile likelihood would not perform well in this situation because profile likelihoods only have the correct asymptotic behavior when used to eliminate a fixed dimensional nuisance parameter (see discussion of Royall, 2000).

However, if repeated observations on each of the i units are collected, we can often construct a conditional likelihood for β that is free of the intercepts. Our true model is now $E(Y_{it}) = \alpha_i + \beta x_{it}$ where $t = 1, \ldots, T$ are the repeated measurements on each of the i units (assume also that the z_i's do not depend on t). Now a conditional likelihood for β may be obtained by conditioning on an appropriate sufficient statistic that accounts for the within unit clustering (Pan, 2002; Neuhaus & Kalbfleisch, 1998; Diggle *et al.*, 2002). Note that β requires careful interpretation in this situation, as the between and within unit (cluster) effects must be distinguished.

Without repeated measurements on each unit, there is not an obvious solution to this problem. Thus we may simply not be able to account for the effect of the missing covariate and hence the object of inference will not be the object of interest. While this may be frustrating, is not surprising.

6 Discussion

When choosing a working model, there are four general cases to consider. They are defined by the answers to two questions: (1) Are the z's important (is their coefficient $\gamma \neq 0$)? and (2) Are the z's correlated with x, the regressor of interest? The first question is about the unknown true model. While the sample data represent empirical evidence about the answer to this question, some uncertainty always remains. By contrast, the second question is about the relationship between x and z, independent of the true model. It can be answered definitively by examining x and z. These four cases correspond to the four cells in a 2×2 table (table 1):

Our reasoning is as follows: including in the working model covariates that are actually unimportant has little effect on the likelihood when those covariates are not correlated with the regressor of interest. (This is true when the number of covariates is small in relation to the number of observations). Including in the working model covariates that contribute nothing to the regression function (i.e., unimportant covariates) but are correlated with the regressor of interest flattens the likelihood (weakens the evidence), but does not distort it. Excluding important covariates from the working model simply weakens the evidence when they are not correlated with the regressor of interest, but

Table 1

Guidelines for including additional covariates (z) in the working model.
** : to decrease the probability of observing weak evidence ('for efficiency')*
***: to control the probability of observing misleading evidence ('for validity').*

	z's **uncorrelated with** *x*	*z*'s **correlated with** *x*
z's **unimportant**	In	Out*
z's **important**	In*	In**

badly distorts the evidence when they are. Thus, if you have a few covariates that are uncorrelated with the regressor of interest, put them in the model. There is little to lose and much to gain.

It is also important to remember that working models chosen according to a 'non-evidential' criteria may not yield a model that reliably represents the statistical evidence about the object of interest. For example, if the goal is to maximize predictive accuracy (in terms of AIC, say), then we saw that important covariates that are correlated with the regressor of interest will likely be eliminated from the model. This renders the evidence about the object of interest strongly misleading if that covariate is indeed important. Although the resulting collection of covariates may be the most predictively accurate, we will not be able to reliably use that working model to learn about the relationships between the covariates and the response.

Of course, the reverse is also true: if a working model is chosen so that it characterizes the statistical evidence correctly, it may not have the highest predictive accuracy or be the most parsimonious. Hence, a comprehensive statistical analysis may require presenting two different working models: one for prediction and one for describing the statistical evidence about the parameters of interest.

A The Probability of Observing Misleading Evidence

For a general treatment, let $\widehat{\beta}_w$ be the working model maximum likelihood estimate of the true regression coefficient for X, which we denote as β_t $(t, w = l, s$ for large or small model as the case may be). Represent its expected value under the true model by $E_t[\widehat{\beta}_w] = \beta_{w|t}$, variance under the working model by $\mathrm{Var}_w[\widehat{\beta}_w] = V_w$ and estimated variance by $\widehat{V}_w$. As the sample size grows, $\widehat{V}_w \rightarrow E_t[\widehat{V}_w] = V_{w|t}$ by the law of large numbers.

Now the working model likelihood ratio for $H_1 : \beta = \beta_1$ over $H_2 : \beta = \beta_2$ is approximately

$$\frac{L_w(\beta_1)}{L_w(\beta_2)} = \frac{\exp\{-\frac{1}{2V_{w|t}}(\widehat{\beta}_w - \beta_1)^2\}}{\exp\{-\frac{1}{2V_{w|t}}(\widehat{\beta}_w - \beta_2)^2\}} = \exp\left\{-\frac{1}{2V_{w|t}}[\beta_1^2 - \beta_2^2 + 2\widehat{\beta}_w(\beta_2 - \beta_1)]\right\}.$$

And the probability of observing a working likelihood ratio larger than k is

$$P\left(\frac{L_w(\beta_1)}{L_w(\beta_2)} \geq k\right) = \begin{cases} P\left(\widehat{\beta}_w > \frac{V_{w|t}\ln k}{(\beta_1-\beta_2)} + \frac{\beta_1+\beta_2}{2}\right) & \beta_1 - \beta_2 > 0 \\ \\ P\left(\widehat{\beta}_w < \frac{V_{w|t}\ln k}{(\beta_1-\beta_2)} + \frac{\beta_1+\beta_2}{2}\right) & \beta_1 - \beta_2 < 0 \end{cases}.$$

But under the true model we have that $Z = (\widehat{\beta}_w - \beta_{w|t})/\sqrt{V_{w|t}}$ is approximately normally distributed yielding

$$P_t\left(\frac{L_w(\beta_1)}{L_w(\beta_2)} \geq k\right) = \begin{cases} \Phi\left[-\frac{\ln k\sqrt{V_{w|t}}}{(\beta_1-\beta_2)} - \frac{\beta_1+\beta_2}{2\sqrt{V_{w|t}}} + \frac{\beta_{w|t}}{\sqrt{V_{w|t}}}\right] & \beta_1 - \beta_2 > 0 \\ \\ \Phi\left[\frac{\ln k\sqrt{V_{w|t}}}{(\beta_1-\beta_2)} + \frac{\beta_1+\beta_2}{2\sqrt{V_{w|t}}} - \frac{\beta_{w|t}}{\sqrt{V_{w|t}}}\right] & \beta_1 - \beta_2 < 0 \end{cases}$$

where $\Phi[\cdot]$ is the standard normal distribution function.

Now we can always write $E_t[\widehat{\beta}_w] = \beta_{w|t} = \beta_t + K^*$, so the expectation is equal to the object of interest plus bias (the bias term K^* may be zero). It follows that the probability of observing misleading evidence under the working model is

$$P_t\left(\frac{L_w(\beta)}{L_w(\beta_t)} \geq k\right) = \Phi\left[-\frac{\sqrt{V_{w|t}}\ln k}{|\beta - \beta_t|} - \frac{|\beta - \beta_t|}{2\sqrt{V_{w|t}}} + \text{sign}(\beta - \beta_t)\frac{B}{\sqrt{V_{w|t}}}\right]. \qquad (9)$$

Now let the alternative approach the true value with $\beta - \beta_t = c\sqrt{\text{Var}_t[\widehat{\beta_t}]} = c\sqrt{V_{t|t}}$. The bias must also be scaled in the same fashion, so $K^* = c_K\sqrt{V_{t|t}}$. Then we have

$$P_t\left(\frac{L_w(\beta_1)}{L_w(\beta_t)} \geq k\right) = \Phi\left[-\frac{q\ln k}{|c|} - \frac{|c|}{2q} + \text{sign}(c)\frac{c_k}{q}\right] \qquad (10)$$

where $q = \sqrt{\text{Var}_t[\widehat{\beta}_w]/\text{Var}_t[\widehat{\beta_t}]} = \sqrt{V_{w|t}/V_{t|t}}$ is the ratio of variances between the two different estimators for β.

Acknowledgement

The author is grateful to Richard Royall for sharing his thoughts and some empirical studies on this subject, to Wesley Eddings for comments on an earlier draft, and to a reviewer for very detailed comments that improved this paper. This work was presented at the 2002 IMS meeting in Banff, CA.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory.* Eds. B.N. Petrov and F. Csaki. Budapest: Akademiai Kiado.

Barnard, G.A. (1947). "Review of Wald, A. *Sequential Analysis*". *Journal of the American Statistical Association*, **42**, 658–664.

Berger, J.O. & Wolpert, R.L. (1988). *The likelihood principle.* Hayward, California: Institute of Mathematical Statistics.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, **53**, 259–326.

Blume, J.D. (2002). Likelihood methods for measuring statistical evidence. *Statistics in Medicine*, **21**, 2563–2599.

Burnham, K.P. & Anderson, D.R. (1998). *Model selection and inference: A practical information-theoretic approach.* New York: Springer-verlag.

Diggle, P.J., Heagerty, P., Liang, K.Y. & Zeger, S.L. (2002). *Analysis of longitudinal data (2nd ed).* Oxford: Oxford University Press.

Hacking, I. (1965). *Logic of Statistical Inference.* New York: Cambridge University Press.

Harrell, F.E. (2001). *Regression modeling strategies.* New York: Springer-verlag.

Montgomery, D. & Peck, E. (1992). *Introduction to linear regression analysis.* New York, NY: John Wiley & Sons.

Neter, J., Wasserman, W. & Kutner, M.H. (1990). *Applied linear statistical models (3rd ed).* Boston MA: Irwin.

Neuhaus, J.M. & Kalbfleisch, J.D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, **54**, 638–645.

Pan, W. (2002). A note on the use of marginal likelihood and conditional likelihood in analyzing clustered data. *The American Statistician*, **56**(3), 171–174.

Pratt, J.W. (1977). 'Decisions' as statistical evidence and Birnbaum's confidence concept. *Synthese*, **36**, 59–69.

Royall, R.M. (1997). *Statistical Evidence: A likelihood paradigm.* London: Chapman and Hall.

Royall, R.M. (2000). On the probability of observing misleading statistical evidence (with discussion). *Journal of the American Statistical Association*, **95**(451), 760–767.

Royall, R.M. & Tsou, T.S. (2003). Interpreting statistical evidence using imperfect models: robust adjusted likelihood functions. *Journal of Royal Statistical Society, Series B*, **65**(2), 391–404.

Seber, G.A.F. (1977). *Linear Regression Analysis.* New York: John Wiley & Sons.

Smith, C.A.B. (1953). The Detection of Linkage in Human Genetics. *Journal of the Royal Statistical Society, Series B*, **15**, 153–192.

Woodroofe, M. (1982). On model selection and the arc sine laws. *The Annals of Statistics*, **10**(4), 1182–1194.

Zhang, P. (1992). On the distributional properties of model selection criteria. *Journal of the American Statistical Association*, **87**(419), 732–737.

Résumé

Considérez que l'utilisation d'un rapport de probabilité mesure la force d'évidence statistique pour une hypothèse sur un autre. Le travail récent a montré que quand le modèle est correctement spécifié, le rapport de probabilité se trompe rarement. Mais quand le modèle n'est pas, en trompant l'évidence peut être observé tout fait souvent. Ici nous estimons comment choisir un modèle de rétrogradation travaillant pour que l'évidence statistique soit correctement représentée aussi souvent qu'il serait sous le vrai modèle. Nous soutenons que les critères pour choisir un modèle travaillant devraient être combien de fois il représente correctement l'évidence statistique de l'objet d'intérêt (le coefficient de rétrogradation dans le vrai modèle). Nous voyons que la tromperie de l'évidence de l'objet d'intérêt sera mieux observée quand le modèle travaillant est choisi selon d'autres critères (par ex., la parcimonie ou l'exactitude prophétique).

[*Received March 2004, accepted March 2005*]