



Taylor & Francis
Taylor & Francis Group

A Closer Look at the Deviance

Author(s): Trevor Hastie

Source: *The American Statistician*, Feb., 1987, Vol. 41, No. 1 (Feb., 1987), pp. 16-20

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: http://www.jstor.com/stable/2684312

REFERENCES
Linked references are available on JSTOR for this article:
http://www.jstor.com/stable/2684312?seq=1&cid=pdf-reference#references_tab_contents
You may need to log in to JSTOR to access the linked references.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms



JSTOR

American Statistical Association and *Taylor & Francis, Ltd.* are collaborating with JSTOR to digitize, preserve and extend access to *The American Statistician*

This content downloaded from
171.66.162.169 on Mon, 20 Jul 2020 18:56:31 UTC
All use subject to https://about.jstor.org/terms

A Closer Look at the Deviance

TREVOR HASTIE*

Prediction error and Kullback–Leibler distance provide a useful link between least squares and maximum likelihood estimation. This article is a summary of some existing results, with special reference to the deviance function popular in the GLIM literature.

KEY WORDS: Exponential family; GLIM; Kullback–Leibler distance; Least squares estimation; Maximum likelihood estimation.

1. INTRODUCTION

The residual sum of squares (RSS), least squares estimation (LS), prediction error, and cross-validated RSS are all familiar terms in the ordinary regression analysis of a quantitative or "continuous" response variable. Deviance, maximum likelihood (ML) estimation, Kullback–Leibler (KL) distance, and AIC (Akaike information criterion) are similar concepts when modeling using likelihood principles. In particular the class of GLIM models (generalized linear models) (McCullagh and Nelder 1983; Nelder and Wedderburn 1972) are foremost in mind. For these models the response, Y, is assumed to follow a one-parameter exponential distribution with density

$$f_\theta(y) = \exp[\theta y - \psi(\theta)]\, dG(y), \tag{1}$$

where θ is the natural or canonical parameter, $\psi(\theta)$ is the normalizing function, and G is the carrier measure. The mean is given by $\mu = d\psi(\theta)/d\theta$ and is monotone in θ, and thus the density can be indexed, alternatively, by μ. A common example is the binomial distribution, $Y \sim B(n, p)$, for which $\theta = \log(p/(1-p))$ and $\psi(\theta) = n\ln(1 + \exp(\theta))$. Typically one models the natural parameter θ, or some other function of μ, as a linear model $\boldsymbol{\beta}^t\mathbf{x}$, where $\mathbf{x}$ is a vector of p covariates.

This note describes a framework that unifies these concepts and exposes the many analogies between the two modes of estimation. The ideas presented here are not new; rather they are established results simply stated for readability. Since the common link is the KL distance function, I start off with a definition and brief description of it.

2. THE KULLBACK–LEIBLER DISTANCE

The KL distance (Kullback 1959) measures the discrepancy between two densities, f_{θ_1} and f_{θ_2}, and is defined by

$$K(\theta_1, \theta_2) = 2E_{\theta_1} \ln[f_{\theta_1}(Y)/f_{\theta_2}(Y)], \tag{2}$$

where E_{θ_1} refers to expectation with respect to the density f_{θ_1}. (I have introduced a factor 2 for convenience; the original definition is without it.) For the remainder of the article I will consider the special case in which $f \in F$, the family of one-parameter exponential densities, indexed by the natural parameter θ as in (1): $f_\theta(y) = \exp[\theta y - \psi(\theta)]\, dG(y)$. Since the mean $E_\theta Y = \mu(\theta)$ is monotone in θ, we can, alternatively, index the densities by μ. See Efron (1978a) for a comprehensive discussion on this family of densities. It follows easily from (1) and (2) that

$$K(\mu_1, \mu_2) = 2[(\theta_1 - \theta_2)\mu_1 - (\psi(\theta_1) - \psi(\theta_2))]. \tag{3}$$

One can think of $K(\mu_1, \mu_2)$ as measuring the distance between the means. Intuitively, $K(\mu_1, \mu_2)$ averages a measure of discrepancy between the two density functions over their support and as such gives more weight (in this average) to areas of higher probability, as determined by the density f_{μ_1}. K is asymmetric in its arguments and hence not a distance function in the true sense. However, $K(\mu_1, \mu_2) \geq 0$ with equality iff $f_{\mu_1} \equiv f_{\mu_2}$. A special case of K is when f is the standard Gaussian density, in which case $K(\mu_1, \mu_2) = (\mu_1 - \mu_2)^2$.

We can take this distance interpretation one step further and regard the positive *random variable* $K(Y, \mu)$ as a measure of the deviation of the random variable Y from the mean μ. According to (2), this is given by $K(Y, \mu) = 2E_Y \ln[f_Y(Y^*)/f_\mu(Y^*)]$, where we average over Y^* with density indexed by Y: $f_Y(y^*)$. Hoeffding's representation of the exponential family (see Efron 1978a) uses such a construct:

$$f_\mu(y) = f_y(y)\exp[-\tfrac{1}{2}K(y, \mu)] \tag{4}$$

and thus

$$K(y, \mu) = 2\ln[f_y(y)/f_\mu(y)]. \tag{5}$$

In GLIM circles this quantity is called the deviance (or component of deviance), where y is a realization of Y and thus $K(y, \mu)$ is a realization of $K(Y, \mu)$. In general the deviance between two models μ_1 and μ_2 at observation y is defined as

$$D(\mu_1, \mu_2; y) = 2\ln[f_{\mu_1}(y)/f_{\mu_2}(y)]. \tag{6}$$

Note.

1. If the first argument is also y, then from (5) and (6) $D(y, \mu; y) = K(y, \mu)$, which is usually written as $D(y, \mu)$. This is the deviance between the model μ and the "saturated" model y (i.e., it fits y exactly).
2. $D(\mu_1, \mu_2; Y)$ is a random variable, whereas $K(\mu_1, \mu_2)$ is not; in fact $K(\mu_1, \mu_2) = E_{\mu_1} D(\mu_1, \mu_2; Y)$. $D(\mu_1, \mu_2; y)$ is a realization of $D(\mu_1, \mu_2; Y)$.
3. We are probably more familiar with a deviance for a random sample of iid observations $\mathbf{y}^t = (y_1, y_2, \ldots, y_n)$. If two models for the mean are μ_1 and μ_2, then the KL distance between $\mathbf{y}$ and $\boldsymbol{\mu}_1 = \mu_1\mathbf{1}$ is

*At the time this document was prepared, Trevor Hastie was a biostatistician at the Institute for Biostatistics, South African Medical Research Council. He is currently a member of the research staff at AT&T Bell Labs, Murray Hill, NJ 07974. The author thanks the referees and an associate editor for their constructive comments on an earlier draft of this article.

 © *1987 American Statistical Association*

This content downloaded from
171.66.162.169 on Mon, 20 Jul 2020 18:56:31 UTC
All use subject to https://about.jstor.org/terms

$$K(\mathbf{y}, \boldsymbol{\mu}_1) = 2 \sum_{i=1}^{n} [\ln f_{y_i}(y_i) - \ln f_{\mu_1}(y_i)]$$
$$= D(\mathbf{y}, \boldsymbol{\mu}_1), \tag{7}$$

and

$$D(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \mathbf{y}) = D(\mathbf{y}, \boldsymbol{\mu}_2) - D(\mathbf{y}, \boldsymbol{\mu}_1)$$
$$= K(\mathbf{y}, \boldsymbol{\mu}_2) - K(\mathbf{y}, \boldsymbol{\mu}_1)$$
$$= 2 \sum_{i=1}^{n} [\ln f_{\mu_1}(y_i) - \ln f_{\mu_2}(y_i)]. \tag{8}$$

4. Once again $D(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \mathbf{y}) \neq K(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$, although in this case the deviance is an estimate of the KL distance.

If f is standard Gaussian, then $K(y, \mu) = (y - \mu)^2$, and in general we can regard $K(y, \mu)$ as measuring the prediction error in using μ to predict y; in this context we regard K as our loss function. For a non-Gaussian example of K, consider the binomial distribution

$$K(y/n, \mu) = 2\{y \ln [(y/n)/\mu] + (n - y)\ln[(1-y/n)/(1 - \mu)]\}, \tag{9}$$

or in the binary case ($n = 1$),

$$K(y, \mu) = -2[y \ln \mu + (1 - y)\ln(1 - \mu)]. \tag{10}$$

[See Efron (1978b) for a detailed discussion of this and other measures of prediction error in the binary case.]

3. THE OPTIMAL PREDICTOR

In what follows I will define quantities in the general form and then follow it immediately with its more familiar squared error counterpart. With $K(Y, \eta)$ as our loss function and predictor η, the risk or expected prediction error (EPE) is given by

$$\text{EPE} = E\,K(Y, \eta), \qquad \text{EPE} = E\,(Y - \eta)^2, \tag{11}$$

where the expectations are with respect to the distribution of Y. If $\mu = EY$, then since $K(Y, \eta) = 2 \ln[f_Y(Y)/f_\eta(Y)] = 2 \ln[f_Y(Y)/f_\mu(Y)] + 2 \ln[f_\mu(Y)/f_\eta(Y)]$, we have

$$E\,K(Y, \eta) = E\,K(Y, \mu) + K(\mu, \eta),$$
$$E\,(Y - \eta)^2 = \text{var}\,Y + (\mu - \eta)^2. \tag{12}$$

This immediately gives the familiar result that the mean is the best predictor in terms of mean squared prediction error and, in general, in terms of expected KL prediction error.

4. ESTIMATION

Suppose that we have a random sample $y_1, y_2, \ldots, y_n$ from f_μ and we wish to estimate μ. It is customary in the squared error case to find that η that minimizes $(1/n)\Sigma(y_i - \eta)^2$, an estimate of $E(Y - \eta)^2$. In general, we might minimize

$$1/n \sum_{i=1}^{n} K(y_i, \eta) = 2/n \sum_{i=1}^{n} \ln[f_{y_i}(y_i)/f_\eta(y_i)]$$
$$= D(\mathbf{y}, \eta\mathbf{1})/n, \tag{13}$$

an estimate of $E\,K(Y, \eta)$.

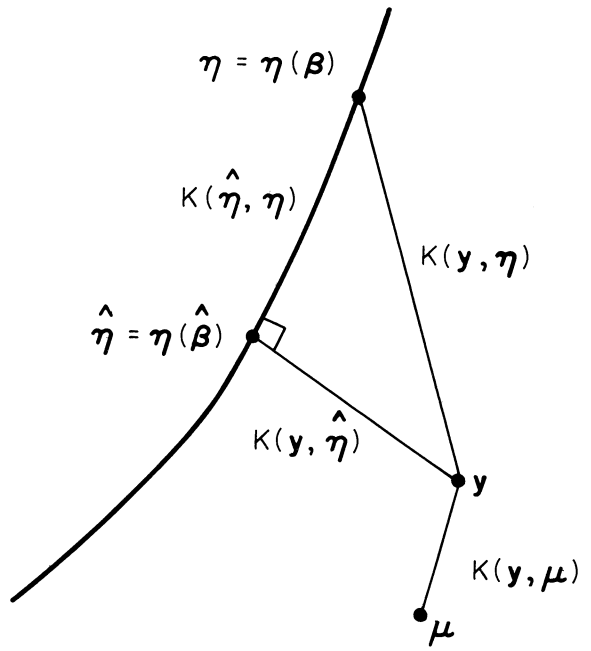


Figure 1. Illustration of the Various KL Distances Between a Density f_μ Indexed by its Mean $\boldsymbol{\mu}$, an Observation Vector $\mathbf{y}$ From f_μ, and a Parameterized Family of Approximating Densities f_η. In this case $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\beta}) = \beta_0\mathbf{1} + \beta\mathbf{x}$. The MLE $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}(\hat{\boldsymbol{\beta}}) = \hat{\beta}_0\mathbf{1} + \hat{\beta}\mathbf{x}$ indexes the member of the family closest in KL distance to $\mathbf{y}$.

From (13) it is clear that this procedure corresponds exactly to ML estimation. Alternatively, up to a factor $1/n$, (13) is the KL distance between the models $f_{\mathbf{y}}$ and $f_{\boldsymbol{\eta}}$, where $\mathbf{y}^t = (y_1, y_2, \ldots, y_n)$ and $\boldsymbol{\eta} = \eta\mathbf{1}$. From this viewpoint ML estimation corresponds to finding the member of the one-parameter family $f_{\boldsymbol{\eta}}$ closest to $\mathbf{y}$. Figure 1 depicts the situation; there is no covariate in this case [see Efron (1978a) for more details of the geometric interpretation].

Estimation With Covariates

If X is a covariate and we wish to model the mean or regression function $E(Y|x) = \mu(x)$ by $\eta(x)$, then (12) becomes

$$E_X E_{Y|X}\,K(Y, \eta(X)) = E_X E_{Y|X}\,K(Y, \mu(X)) + E_X K(\mu(X), \eta(X)),$$
$$E_X E_{Y|X}(Y - \eta(Y))^2 = E_X\,\text{var}(Y|X) + E_X(\mu(X) - \eta(X))^2. \tag{14}$$

(Usually we condition on the observed covariates—here I have averaged over the distribution of X as well; the conclusions are the same.) We once again see that the optimal choice for $\eta(x)$ is $\mu(x)$. Often $\eta(x)$ is constrained; that is, $\eta(x) = \beta_0 + \beta x$ or $\eta(x) = s(x)$, where s is a "smooth" function. In this case minimizing (14) corresponds to finding that member of the constrained family closest to $\mu(x)$ in expected KL distance [see Hastie and Tibshirani (1986) for more details in the case of $s(x)$]. Figure 1 once again illustrates the minimization.

Given a sample $(y_1, x_1), \ldots, (y_n, x_n)$ of iid pairs of (Y, X), we can estimate (14) by

This content downloaded from
171.66.162.169 on Mon, 20 Jul 2020 18:56:31 UTC
All use subject to https://about.jstor.org/terms

$$1/n \sum_{i=1}^{n} K(y_i, \eta(x_i)) = 2/n \sum_{i=1}^{n} \ln[f_{y_i}(y_i)/f_{\eta(x_i)}(y_i)]$$

$$= D(\mathbf{y}, \boldsymbol{\eta}(\mathbf{x}))/n. \quad (15)$$

Once again, minimizing this quantity, the mean deviance, corresponds to ML estimation. As an example consider the logistic regression problem: y_i ($i = 1, n$) are assumed to be independent Bernoulli samples with $E(y_i|x_i) = \mu(x_i)$, and typically we assume the model $\log[\mu(x_i)/(1 - \mu(x_i))] = \text{logit}[\mu(x_i)] = \beta_0 + \beta x_i$. (This corresponds to modeling the natural parameter θ_i mentioned in Section 1.) One can easily show that $D(\mathbf{y}, \boldsymbol{\mu}) = -2\Sigma_i[(\beta_0 + \beta x_i)y_i + \ln(1 - \mu(x_i))]$, which is exactly $-2 \times$ log-likelihood. For other generalized models there will also be, typically, a term depending only on $\mathbf{y}$.

5. BIAS DUE TO OVERFIT

So far we have seen that the regression function $\mu(x) = E(Y|x)$ minimizes the theoretical quatity (14). The most common methods of estimation, ML and LS, can be thought of as using an estimate of this theoretical quantity as a criterion for parameter estimation. Often we do not stop there but wish to try a variety of different models and pick the best one. For example, we might have a number of different linear models, each with a different number of parameters. Alternatively, we might have a variety of nonlinear models, each with a slightly different functional form. In nonparametric regression, the window size or span of the smoother is a parameter that must be chosen. We might think of using the criterion that was minimized in the parameter search *within* a model to select *between* the different models. Since the parameters of each model are chosen to fit the data best, however, it is not surprising that in general the model with the most parameters, or degrees of freedom, wins (fewest degrees of freedom for error). So what has gone wrong? For an entertaining account (with discussion) of how to avoid such overuse of the data, the reader is referred to the papers of Stone (1974, 1977) on cross-validation. I summarize these ideas within the convenient setting of prediction error.

5.1 Bias Due to Overfit: Least Squares Regression

I first examine the LS linear regression model in more detail. Suppose that $y_i = \mu(\mathbf{x}_i) + \varepsilon_i$ ($i = 1, \ldots, n$), ε_i iid, $\mathbf{x}_i$ are vectors of p covariates, and we approximate $\mu(\mathbf{x}_i)$ by the linear model $\eta(\mathbf{x}_i) = \boldsymbol{\beta}'\mathbf{x}_i$. The coefficients are estimated by least squares, and the fitted values are denoted by $\hat{\eta}(\mathbf{x}_i)$. A familiar result is that conditional on the observed covariates,

$$E(\text{RSS})/n = E \sum_{i=1}^{n} (y_i - \hat{\eta}(\mathbf{x}_i))^2/n$$

$$= \sigma^2 + \sum_{i=1}^{n} [\mu(\mathbf{x}_i) - E\hat{\eta}(\mathbf{x}_i)]^2/n$$

$$- [p/n]\sigma^2, \quad (16)$$

where var $\varepsilon_i = \sigma^2$ (see Lemma 1, Appendix). The term $[p/n]\sigma^2$ is the average variance of $\hat{\eta}(\mathbf{x}_i)$. If, however, we are able to draw a fresh sample of y's, say a y_i^* at each $\mathbf{x}_i$, to measure the prediction error of our estimated mean, we find

$$E(\text{RSS}^*/n) = E \sum_{i=1}^{n} (y_i^* - \hat{\eta}(\mathbf{x}_i))^2/n$$

$$= \sigma^2 + \frac{1}{n}\Sigma(\mu(\mathbf{x}_i) - E\hat{\eta}(\mathbf{x}_i))^2 + [p/n]\sigma^2 \quad (17)$$

(see Lemma 2, Appendix). Equation (17) makes a lot of sense: if we know the true regression function, our baseline average squared prediction error will be σ^2. Since the estimated regression function has bias and variance, this must add to the baseline error. From (16) and (17) we see that RSS/n is biased downwards as an estimate of the EPE of $\hat{\eta}$ by the amount $2[p/n]\sigma^2$.

In a more general setting, suppose that we use a linear smoother $s(x)$ (see, e.g., Cleveland 1979) to estimate the regression function of y on a single covariate x. Let $\mathbf{s} = S\mathbf{y}$ denote the vector of n fitted values, where S is the $n \times n$ smoother matrix. We find

$$E(\text{RSS}/n) = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} \text{bias}_i^2$$

$$- \frac{1}{n}\text{tr}(2S - S'S)\sigma^2, \quad (18)$$

whereas

$$E(\text{RSS}^*/n) = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} \text{bias}_i^2 + \frac{1}{n}\text{tr}(S'S)\sigma^2 \quad (19)$$

(Rice 1984; Tibshirani 1984). Both the trace terms in Equations (18) and (19) have been used to define the "degrees of freedom" of a linear smoother S, as has tr(S) (Cleveland 1979; Cleveland and Devlin 1986). The difference between (19) and (18) is $(1/n)\ 2\ \text{tr}(S)\sigma^2$, and in general the bias is of the form $(1/n)\ 2p\sigma^2$, where p represents the "degrees of freedom" of the fit. This has led to a variety of corrections, such as C_p (Mallows 1973), cross-validation (Stone 1974, 1977), and others, which attempt to correct the RSS for this bias:

1. C_p adds a term $2p\hat{\sigma}_e^2/n$ to RSS/n, where $\hat{\sigma}_e^2$ is an estimate of σ^2 obtained from a "full" model.
2. Cross-validation replaces $\hat{\eta}(\mathbf{x}_i)$ by $\hat{\eta}_{(i)}(\mathbf{x}_i)$, where the latter is an estimate using the same "prescription" as $\hat{\eta}$ but with the ith point removed. This fix in fact highlights the problem in using RSS: $\hat{\eta}(\mathbf{x}_i)$ is correlated with y_i, since it is a linear function of y_i, and as such will tend to reproduce any random anomolies present in y_i; $\hat{\eta}_{(i)}(\mathbf{x}_i)$ is not a function of y_i. See Atkinson (1981) for a full account of such corrections and Efron (1985) for a discussion on the relative merits of the various corrections in this case and those pertaining to the next section.

5.2 Bias Due to Overfit: The General Case

We have seen that the KL distance and deviance measure prediction error in the exponential family; in the general case does bias due to overfit occur, and in the same form? For the class of generalized linear models, suppose that $\hat{\eta}_i = \mathbf{b}'\mathbf{x}_i$ is the ML estimate of the linear predictor η_i. For the logistic regression problem $\text{logit}(\mu(\mathbf{x})) = \eta(\mathbf{x})$. Let

This content downloaded from
171.66.162.169 on Mon, 20 Jul 2020 18:56:31 UTC
All use subject to https://about.jstor.org/terms

$\hat{\mu}(\mathbf{x}_i)$ denote the fitted mean value corresponding to $\hat{\eta}(\mathbf{x}_i)$, and let $\gamma(\mathbf{x}_i)$ be the fitted value corresponding to $E(\hat{\eta}(\mathbf{x}_i))$. Then it can be shown that

$$E \sum_{i=1}^{n} K(y_i, \hat{\mu}(\mathbf{x}_i))/n$$
$$= E\, K(\mathbf{y}, \hat{\boldsymbol{\mu}})/n$$
$$= E\, D(\mathbf{y}, \hat{\boldsymbol{\mu}})/n$$
$$= \sum_{i=1}^{n} E\, K(y_i, \mu(\mathbf{x}_i))/n + \Sigma\, K(\mu(\mathbf{x}_i), \gamma(\mathbf{x}_i))/n$$
$$- \sum_{i=1}^{n} E\, K(\hat{\mu}(\mathbf{x}_i), \gamma(\mathbf{x}_i))/n \qquad (20)$$

(Efron 1986; Tibshirani 1984). The E in (20) refers to conditional expectation, given the values of $\mathbf{x}_i$. The first term on the right of (20) corresponds to σ^2, the second to the average squared bias, and the third is a variance type term (all measured using the KL distance as a loss function). In addition, it can be shown that the last term is approximately p/n, where p is the dimension of the subspace spanned by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Tibshirani (1984) showed that the decomposition (20) is approximately valid for more general methods of modeling, such as local-likelihood smoothing.

On the other hand, given a test sample y_i^* $(i = 1, n)$, we have

$$E \sum_{i=1}^{n} K(y_i^*, \hat{\mu}(\mathbf{x}_i))/n$$
$$= E\, K(\mathbf{y}^*, \hat{\boldsymbol{\mu}})/n$$
$$= \sum_{i=1}^{n} E\, K(y_i^*, \mu(\mathbf{x}_i))/n + \Sigma\, K(\mu(\mathbf{x}_i), \gamma(\mathbf{x}_i))/n$$
$$+ \sum_{i=1}^{n} E\, K(\gamma(\mathbf{x}_i), \hat{\mu}(\mathbf{x}_i))/n. \qquad (21)$$

The last term is also a variance of fit type term, although not exactly the same as in (20) because of the asymmetry of the KL distance. It does, however, have expected value about p/n. So the analogy goes through and outlines the argument behind the AIC correction factor (Akaike 1973; Atkinson 1981), which, like C_p, attempts to correct the mean deviance for bias due to overfit:

$$\text{AIC}/n = \text{Deviance}/n + 2p\hat{\sigma}^2/n \qquad (22)$$

(in some cases, such as the binomial and Poisson, $\sigma^2 = 1$).

5.3 Example

As a straightforward illustration let us examine again the Bernoulli distribution. Given a sample y_i $(i = 1, \ldots, n)$, where $y_i = 0$ or 1, suppose that $\hat{\mu} = \Sigma y_i/n$ and $E_{y_i} = \mu$. Then

$$E_{Y^*} \sum_{i=1}^{n} K(y_i^*, \hat{\mu})/n$$
$$= -2[\mu \ln \hat{\mu} + (1 - \mu)\ln(1 - \hat{\mu})].$$

However, using a Taylor expansion,

$$\sum_{i=1}^{n} K(y_i, \hat{\mu})/n = -2[\hat{\mu} \ln \hat{\mu} + (1 - \hat{\mu})\ln(1 - \hat{\mu})]$$
$$= -2[\mu \ln \hat{\mu} + (1 - \mu)\ln(1 - \hat{\mu})]$$
$$- 2\frac{(\hat{\mu} - \mu)^2}{\mu(1 - \mu)} + R,$$

where $E(R) = O(1/n^2)$. Thus

$$E \sum_{i=1}^{n} K(y_i^*, \hat{\mu})/n$$
$$= E \sum_{i=1}^{n} K(y_i, \hat{\mu})/n - 2/n + O(1/n^2).$$

Note that in this case $p = 1$.

6. ADDITIVITY

Finally, I draw further attention to the Pythagorean property of both KL distance and squared error (Fig. 2). Suppose that L_1 is a p dimensional linear subspace of R^n, and $\hat{\boldsymbol{\mu}}_1$ is the LS estimate or projection of $\mathbf{y}$ onto L_1. If $\mathbf{f}$ is any vector in L_1, then

$$\| \mathbf{y} - \mathbf{f} \|^2 = \| \mathbf{y} - \hat{\boldsymbol{\mu}}_1 \|^2 + \| \hat{\boldsymbol{\mu}}_1 - \mathbf{f} \|^2, \qquad (23)$$

where $\| \cdot \|$ represents the usual Euclidean norm. In particular, if $\mathbf{f} = \hat{\boldsymbol{\mu}}_2 \in L_2 \subset L_1$ is the projection of $\mathbf{y}$ onto a $q < p$ dimensional subspace of L_1, then

$$\| \mathbf{y} - \hat{\boldsymbol{\mu}}_2 \|^2 = \| \mathbf{y} - \hat{\boldsymbol{\mu}}_1 \|^2 + \| \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2 \|^2. \qquad (24)$$

This result is also true for generalized linear models (Simon 1973) if projection is in the KL distance sense, and we

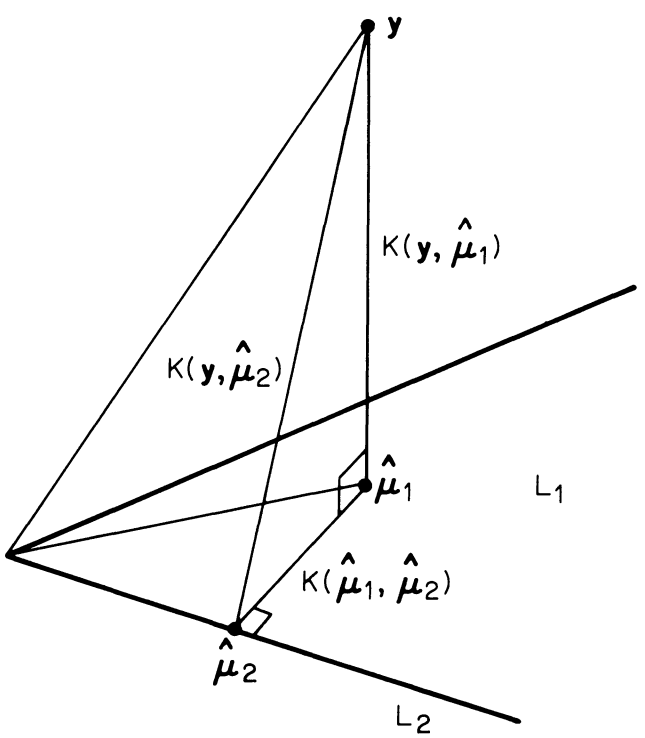


Figure 2. The Pythagorean Property of MLE's. $L_2 \subset L_1$ are linear subspaces of R^n, and $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ are the "projections" in the KL sense, that is, MLE's.

This content downloaded from
171.66.162.169 on Mon, 20 Jul 2020 18:56:31 UTC
All use subject to https://about.jstor.org/terms

model $\boldsymbol{\eta} = \mathbf{x}\boldsymbol{\beta}$, where $\boldsymbol{\eta}$ is the natural parameter and $\mathbf{y}$ the observed response vector. Thus

$$K(\mathbf{y}, \hat{\boldsymbol{\mu}}_2) = K(\mathbf{y}, \hat{\boldsymbol{\mu}}_1) + K(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2). \tag{25}$$

Note:

1. This result is always true if K is replaced with D. Simon's theorem tells us that in this special case of nested linear models $D(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2; \mathbf{y}) = K(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2)$.
2. From 4.1, projection in the KL sense means that $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ are both maximum likelihood estimators (MLE's).
3. A special case of Simon's theorem is used in the derivation of (20): $K(\boldsymbol{\mu}, \boldsymbol{\gamma}) = K(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) + K(\hat{\boldsymbol{\mu}}, \boldsymbol{\gamma})$.

If the true mean $\boldsymbol{\mu}(\mathbf{x}) = E(\mathbf{y}|\mathbf{x})$ is such that $\boldsymbol{\eta}_{\boldsymbol{\mu}(\mathbf{x})} \in L_2$, then Wilks's theorem gives us the familiar result that $K(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2) \sim \sigma^2 \chi^2_{p-q}$.

APPENDIX: PROOF OF DECOMPOSITIONS IN SECTION 5.1

Lemma 1.

$$E(\mathrm{RSS})/n = \sigma^2 + \sum_{i=1}^{n} [\mu(\mathbf{x}_i) - E\hat{\eta}(\mathbf{x}_i)]^2/n - [p/n]\sigma^2.$$

Proof. In vector notation,

$$\mathbf{y} - \hat{\boldsymbol{\eta}} = (\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})(\boldsymbol{\mu} + \boldsymbol{\varepsilon})$$

and

$$\begin{aligned}\mathrm{RSS} &= (\mathbf{y} - \hat{\boldsymbol{\eta}})^t(\mathbf{y} - \hat{\boldsymbol{\eta}}) \\ &= \boldsymbol{\mu}^t(\mathbf{I} - \mathbf{H})\boldsymbol{\mu} + 2\boldsymbol{\mu}^t(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}, \\ &\quad + \boldsymbol{\varepsilon}^t(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon},\end{aligned}$$

where $\mathbf{H}$ is the LS hat matrix. Now $\mathbf{H}\boldsymbol{\mu} = E\hat{\boldsymbol{\eta}}$, the middle term has expectation zero, and the last term has expectation $\sigma^2 \operatorname{tr}(\mathbf{I} - \mathbf{H}) = \sigma^2(n - p)$. The result follows.

Lemma 2.

$$\begin{aligned}E(\mathrm{RSS}^*/n) &= \sigma^2 + \Sigma[\mu(\mathbf{x}_i) - E\hat{\eta}(\mathbf{x}_i)]^2 \\ &\quad + [p/n]\sigma^2.\end{aligned}$$

Proof. Let μ_i denote $\mu(\mathbf{x}_i)$, and so on. Then

$$\begin{aligned}y_i^* - \hat{\eta}_i &= (y_i^* - \mu_i) + (\mu_i - E(\hat{\eta}_i)) \\ &\quad + (E(\hat{\eta}_i) - \hat{\eta}_i).\end{aligned}$$

Since $\hat{\eta}_i$ depends on $\mathbf{y}$ and is independent of $\mathbf{y}^*$, the first and last terms are independent with mean zero and the middle term is not random. The result follows by squaring the left side and summing the terms. $\Sigma_i E(\hat{\eta}_i - E(\hat{\eta}_i))^2$ is the sum of the variance of the fits, and, as in Lemma 1, this is $p\sigma^2$. Notice that such a derivation would not have worked in Lemma 1, since there the first and last term would not be independent.

[*Received January 1986. Revised May 1986.*]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Entropy Maximization Principle," in *Proceedings of the 2nd International Symposium on Information Theory*, pp. 267–281.

Atkinson, A. C. (1981), "Likelihood Ratios, Posterior Odds and Information Criteria," *Journal of Econometrics*, 16, 15–20.

Cleveland, W. (1979), "Robust Locally Weighted Regression and Smoothing of Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Cleveland, W., and Devlin, S. (1986), "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," technical memorandum, Murray Hill, NJ: AT&T Bell Laboratories.

Efron, B. (1978a), "The Geometry of Exponential Families," *The Annals of Statistics*, 6, 362–376.

——— (1978b), "Regression and ANOVA With Zero-One Data: Measures of Residual Variation," *Journal of the American Statistical Association*, 73, 113–121.

——— (1986), "How Biased Is the Apparent Error Rate of a Prediction Rule?," *Journal of the American Statistical Association*, 81, 461–470.

Hastie, T., and Tibshirani, R. (1986), "Generalized Additive Models" (with discussion), *Statistical Science*, 1, 297–318.

Kullback, S. (1959), *Information Theory and Statistics*, New York: John Wiley. [Reprinted (1968), New York: Dover Publications.]

Mallows, C. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

McCullagh, P., and Nelder, J. (1983), *Generalized Linear Models*, London: Chapman & Hall.

Nelder, J., and Wedderburn, R. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Ser. A, 135, 370–384.

Rice, J. (1984), "Bandwidth Choice for Nonparametric Regression," *The Annals of Statistics*, 12, 1215–1230.

Simon, G. (1973), "Additivity of Information in Exponential Family Probability Laws," *Journal of the American Statistical Association*, 68, 478–482.

Stone, M. (1974), "Cross-validatory Choice and Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

——— (1977), "An Asympotic Equivalence of Choice or Model by Cross-validation and Akaike's Criterion," *Journal of the Royal Statistical Society*, Ser. B, 7, 44–47.

Tibshirani, R. (1984), "Local Likelihood Estimation," unpublished Ph.D. thesis, Stanford University.

This content downloaded from
171.66.162.169 on Mon, 20 Jul 2020 18:56:31 UTC
All use subject to https://about.jstor.org/terms