General Prediction Theory and the Role of R^2

Ronald Christensen
Department of Mathematics and Statistics
University of New Mexico
Albuquerque, NM 87131

April 13, 2007

Abstract

We consider the role of R^2 in general prediction models. These include linear models, generalized linear models, nonlinear regression models, nonparametric regression models, and various complex computer models. For sample data y_i and corresponding predictors $\tilde{y}_i$, we suggest that their squared sample correlation, called the squared predictive correlation, is the appropriate measure of the overall predictive ability of the model. For linear models, this gives the standard definition of R^2. The rationale for this suggestion comes from putting the problem into a framework of general prediction theory. In general prediction theory, the usual R^2 from linear models can be viewed as an estimate of the squared multiple correlation coefficient, which is a fundamental measure of how well the best linear predictor works. We extend the idea of the squared multiple correlation coefficient by developing the maximum squared predictive correlation which is a strictly analogous fundamental measure of how well the best predictor works. We then argue that for nonlinear problems, R^2 should be developed as an estimate of the squared predictive correlation. The fact that we are estimating a population parameter distinguishes this from many other attempts at extending R^2 to nonlinear models and also provides a basis for using R^2 to compare nonnested models. Of course, for standard linear models (with an intercept) this generalized procedure gives the standard results. We observe that based on general prediction theory, the actual predictions may be improved by finding new predictors $\hat{y}_i$ obtained by regressing the y_is on the $\tilde{y}_i$s, however we also argue that in most commonly used statistical procedures there is little to be gained by linearization. While R^2 is a fundamental measure of predictive ability, it is not a direct measure of goodness of fit. Nonetheless, the relative sizes of R^2 do bear on the issue of goodness of fit through the concept of linearization. While most of these ideas are highly intuitive, we provide a basis for them in general prediction theory, revisit the issue of R^2 for regression through the origin, and examine estimation and residual plots.

KEY WORDS: Best Linear Prediction; Best Prediction; Generalized Linear Models; Linear Models; Multiple Correlation Coefficient; Nonlinear Regression; Nonparametric Regression.

1. Introduction

The coefficient of determination R^2 is a useful measure in linear models. There have been numerous attempts to extend the idea of R^2 to nonlinear models, often creating measures that mimic the behavior of the linear models R^2, see Kvalseth (1985). Our approach is based on predicting an observable random variable y using an observable random vector x. A predictor $f(x)$ is evaluated using squared error prediction loss, $\mathrm{E}[y - f(x)]^2$. Lack of fit and goodness of fit relate to how far or close a (linear) predictor comes to the best (linear) predictor.

In a (correct) linear model, R^2 estimates the squared multiple correlation coefficient, that is, the maximum of the squared correlation coefficients between y and linear predictors based on x. It is also the squared correlation between y and the best linear predictor. In more general problems, we suggest that R^2 should estimate the squared correlation between y and the best predictor, or the maximum of the squared correlation coefficients between y and arbitrary predictors.

R^2 is often thought to indicate how well a model fits data. There is little to justify such a claim. The size of R^2 is not directly related to lack or goodness of fit. Large R^2 values can occur with demonstrably inadequate models and small R^2 values can occur with perfect models.

EXAMPLE: Table 1 contains Hooker's data on the relationship between atmospheric pressure and the boiling point of water as discussed in Weisberg (1985, p. 28) and Christensen (1996, p. 191). A simple linear regression of pressure on temperature gives $R^2 = (.99588)^2$. Figure 1 is a plot of the residuals versus the predicted values. It shows a palpable lack of fit. In Section 7 we argue that an alternative residual plot is more relevant.

One can also simulate data from a known linear model (using the best predictor), and watch R^2 decrease to 0 as the error variance increases. The absolute size of R^2 is not related to goodness of fit but when comparing alternative models based on the same group of predictors, *relative* sizes of R^2 indicate relative goodness of fit. Based on its theoretical genesis, R^2 is properly considered an internal measure of the predictive ability of the model. (An internal measure estimates the squared correlation with the same data used to fit the model.)

Given y predictors of y are simply functions of x. Suppressing the functional notation, we often denote a predictor as $\tilde{y}$ or $\hat{y}$, with $\hat{y}$ generally used as a modification of the predictor $\tilde{y}$. In practice, a predictor $f(x)$ is often modeled as a member of a family of functions for which data are used to identify (estimate) a particular member of the family denoted $\hat{f}(x)$.

For y and a predictor $\tilde{y}$, the squared correlation is an appropriate theoretical measure of the predictive ability of the model, here called the *squared predictive correlation*. For data y_i, $i = 1, \ldots, n$, and corresponding predictions $\tilde{y}_i$, estimate the squared predictive correlation using

$$R^2 \equiv \frac{(s_{y\tilde{y}})^2}{s_y^2 s_{\tilde{y}}^2} \equiv \frac{(s_{y\tilde{y}})^2}{s_{yy} s_{\tilde{y}\tilde{y}}} = \frac{[\sum_{i=1}^n (y_i - \bar{y}_\cdot)(\tilde{y}_i - \bar{y}^*_\cdot)]^2}{\sum_{i=1}^n (y_i - \bar{y}_\cdot)^2 \sum_{i=1}^n (\tilde{y}_i - \bar{y}^*_\cdot)^2},$$

where $\bar{y}_\cdot^*$ is the mean of the $\tilde{y}_i$s. In the Appendix we show that for a linear model with an intercept this gives the commonly used definition of R^2 which is the sum of squares for regression divided by the sum of squares total. This measure has also been discussed by Mittlbock and Schemper (1995) and notably by Zheng and Agresti (2000), who look at it in the context of generalized linear models rather than general prediction theory. General prediction theory allows us to give some different results and insights. Additional historical perspective is presented in the closing section of this article.

Section 2 shows that the squared correlation between y and the best predictor maximizes the squared correlation between y and arbitrary (nonlinear) predictors. Section 2 also investigates linearized predictors and provides a link between squared predictive correlation and goodness of fit. Predictors are never harmed by linearizing them and, for linearized predictors, the larger the squared predictive correlation, the smaller the expected squared prediction error.

One useful aspect of this theory is that it allows comparison of nonnested models. The best predictor is unknown, and various models for it can be proposed. In binomial regression one might consider both logistic and probit models.

EXAMPLE: Christensen (1997, Table 2.1) reproduces the *Challenger* space shuttle O-ring data. We predict the event that one or more O-rings failed as a function of temperature. The squared predictive correlation for the logit and probit models are $(.58813)^2$ and $(.58201)^2$, respectively, so the logit model looks slightly better although there is not much difference in predictive ability. If we use a quadratic model in temperature, the squared predictive correlations are $(.62177)^2$ and $(.61787)^2$, respectively. Not surprisingly, R^2 goes up when adding a predictor variable.

EXAMPLE: Two standard approaches to dealing with the lack of fit in the Hooker data are 1) to fit a quadratic and 2) regressing log of pressure on temperature. Fitting these models give R^2s of $(.99922)^2$ and $(.99898)^2$, respectively. The values are not comparable, being based on different dependent variables. Exponentiating the fitted values from log pressure on temperature and computing their squared correlation with pressure gives $(.99913)^2$, which is comparable to the R^2 for fitting a quadratic to pressure. These R^2 values are numerical summaries of Figures 2 and 3 which plot y versus the predicted values for the quadratic model and the exponentiated fitted values for log pressure, respectively. Clearly, the two models predict almost equally well.

The idea that a larger R^2 means a better model is complicated by estimating, first the parameters of the prediction models, and then the squared predictive correlations. The double use of the data leads to complications as in linear regression where, because of the estimation process, models with lower R^2 and lower numbers of predictors are often better than models with higher R^2s and more predictors. Theoretically, the best predictor based on p variables can never be better than the best predictor based on those p variables plus additional variables. The worst thing that could happen is that the additional variables are irrelevant. Nonetheless, it is better to drop marginally important variables than to

estimate their parameters even though, outside of linear models, there is no guarantee that the estimate of the squared predictive correlation will increase with increased model size.

Squared predictive correlation assumes only the existence of second moments and conditional expectations, so it applies unchanged to almost any prediction problem. Even in an extreme case such as y taking on only the values 0 and 1, squared error prediction loss has a long history of use in the form of Brier scores, cf. Blattenberger and Lad (1985) and Schmid and Griffith (1998).

Section 2 argues that for nonlinear models R^2 should be defined so as to estimate the squared predictive correlation. The question of estimation arises in two distinct forms: first, as an effort to approximate the theoretically optimal predictor, and second, as an effort to approximate the squared predictive correlation between y and the best predictor. Even if one knows the best predictor, there remains the issue of estimating its squared predictive correlation. In linear models, least squares estimates can be viewed as natural estimates of the best linear predictor and the standard definition of R^2 is a natural estimate of the squared multiple correlation coefficient. Section 3 addresses the issue of estimating the squared predictive correlation for general prediction models (linear or nonlinear), leaving the issue of estimating the best predictor to Section 5.

Section 4 examines the specific problem of defining R^2 for regression through the origin and provides a new measure of R^2 based on general prediction theory.

Section 5 examines method of moments estimation. Estimating equations provide a natural method for finding method of moments estimators in nonlinear prediction problems. Section 6 discusses error estimation and Section 7 gives a result from general prediction theory that provides a theoretical basis for looking at residual plots as a method of detecting lack of fit. Section 8 contains discussion and conclusions.

2. General Prediction Theory

Suppose we have random variables $y, x_1, x_2, \ldots, x_p$. Regression is the problem of predicting y from the values of $x_1, \ldots, x_p$. Let x be the vector $x = (x_1, \ldots, x_p)'$. The best predictor of y is a function $f(x)$ that minimizes the mean squared error, $\mathrm{E}[y - f(x)]^2$. For proofs of Theorems 1 and 2, see Christensen (2002, Sec. 6.3). Other results are proven in the Appendix.

Theorem 1. Let $m(x) \equiv \mathrm{E}(y|x)$, then for any other predictor $f(x)$, $\mathrm{E}[y - f(x)]^2 = \mathrm{E}[y - m(x)]^2 + \mathrm{E}[m(x) - f(x)]^2$; thus $m(x)$ is the best predictor of y.

Technically, the conditional expectation is defined only up to sets of measure 0, but that need not concern us in any of our calculations.

Let $\mathrm{E}(y) = \mu_y$, $\mathrm{Var}(y) = \sigma_{yy}$, $\mathrm{E}(x) = \mu_x$, $\mathrm{Cov}(x) = \Sigma_{xx}$, and $\mathrm{Cov}(x, y) = \Sigma_{xy} = \Sigma'_{yx} = \mathrm{Cov}(y, x)'$. Let β_* be a solution to $\Sigma_{xx}\beta = \Sigma_{xy}$. While it is not crucial, we will assume that Σ_{xx} is nonsingular.

Theorem 2. $\hat{E}(y|x) \equiv \mu_y + (x-\mu_x)'\beta_*$ is the best linear predictor of y and $\mathrm{E}[y-\alpha-x'\beta]^2 = \mathrm{E}[y - \hat{E}(y|x)]^2 + \mathrm{E}[\hat{E}(y|x) - \alpha - x'\beta]^2$.

Consider an arbitrary predictor $\tilde{y}(x)$. This is a function of x alone and not a function of y. Let $\mathrm{E}[\tilde{y}(x)] = \mu_{\tilde{y}}$, $\mathrm{Var}[\tilde{y}(x)] = \sigma_{\tilde{y}\tilde{y}}$, and $\mathrm{Cov}[y, \tilde{y}(x)] = \sigma_{y\tilde{y}}$ with similar notations for other functions of x, e.g., $\sigma_{ym} = \mathrm{Cov}[y, m(x)]$. The *squared predictive correlation* of $\tilde{y}(x)$ is $\mathrm{Corr}^2[y, \tilde{y}(x)]$. The highest squared predictive correlation is obtained by using the best predictor. Note that in the special case where $m(x)$ is the best linear predictor, the highest squared predictive correlation equals the squared multiple correlation coefficient.

Proposition 3. $\mathrm{Cov}[y, \tilde{y}(x)] = \mathrm{Cov}[m(x), \tilde{y}(x)]$, so $\mathrm{Cov}[y, m(x)] = \mathrm{Var}[m(x)] = \sigma_{mm}$ and $\mathrm{Corr}^2[y, m(x)] = \sigma_{mm}/\sigma_{yy}$.

Theorem 4. $\mathrm{Corr}^2[y, \tilde{y}(x)] \leq \mathrm{Corr}^2[y, m(x)]$.

Theorem 4 is also established in Rao (1973, Sec. 4g.1). From Theorem 4, the best regression function $m(x)$ has the highest squared predictive correlation. When we have perfect prediction, the highest squared predictive correlation is 1. In other words, if the conditional variance of y given x is 0, then $y = m(x)$ a.s., and the highest squared predictive correlation is the correlation of $m(x)$ with itself, which is 1. On the other hand, if there is no regression relationship, i.e., if $m(x) = \mu_y$ a.s., then $\sigma_{mm} = 0$, and the highest squared predictive correlation is 0.

We would now like to show that as the squared predictive correlation increases, we get increasingly better prediction. First we need to deal with the fact that high squared predictive correlations can be achieved by bad predictors. Just because $\tilde{y}(x)$ is highly correlated with y does not mean that $\tilde{y}(x)$ is actually close to y. Recall that $\tilde{y}(x)$ is simply a random *variable* that is being used to predict y. As such, $\tilde{y}(x)$ acts as a linear predictor of y, that is, $\tilde{y}(x) = 0 + 1\tilde{y}(x)$. We can apply Theorem 2 to this random variable to obtain a linear predictor that is at least as good as $\tilde{y}(x)$, namely

$$\hat{y}(x) = \mu_y + \frac{\sigma_{y\tilde{y}}}{\sigma_{\tilde{y}\tilde{y}}}[\tilde{y}(x) - \mu_{\tilde{y}}].$$

We refer to such predictors as *linearized predictors*. Note that $\mathrm{E}[\hat{y}(x)] \equiv \mu_{\hat{y}} = \mu_y$,

$$\sigma_{\hat{y}\hat{y}} \equiv \mathrm{Var}[\hat{y}(x)] = \left(\frac{\sigma_{y\tilde{y}}}{\sigma_{\tilde{y}\tilde{y}}}\right)^2 \sigma_{\tilde{y}\tilde{y}} = \frac{(\sigma_{y\tilde{y}})^2}{\sigma_{\tilde{y}\tilde{y}}},$$

and

$$\sigma_{y\hat{y}} \equiv \mathrm{Cov}[y, \hat{y}(x)] = \frac{\sigma_{y\tilde{y}}}{\sigma_{\tilde{y}\tilde{y}}}\sigma_{y\tilde{y}} = \frac{(\sigma_{y\tilde{y}})^2}{\sigma_{\tilde{y}\tilde{y}}}.$$

In particular, $\sigma_{\hat{y}\hat{y}} = \sigma_{y\hat{y}}$, so the squared predictive correlation of $\hat{y}(x)$ is

$$\mathrm{Corr}^2[y, \hat{y}(x)] = \frac{\sigma_{\hat{y}\hat{y}}}{\sigma_{yy}}.$$

In addition, the direct measure of the goodness of prediction for $\hat{y}(x)$ is

$$\mathrm{E}[y - \hat{y}(x)]^2 = \sigma_{yy} - 2\sigma_{y\hat{y}} + \sigma_{\hat{y}\hat{y}} = \sigma_{yy} - \sigma_{\hat{y}\hat{y}}.$$

This leads directly to

Theorem 5. For two linearized predictors $\hat{y}_1(x)$ and $\hat{y}_2(x)$, the squared predictive correlation of $y_2(x)$ is higher if and only if $y_2(x)$ is a better predictor.

It should be noted that linearizing $m(x)$ simply returns $m(x)$. In their discussion, Zheng and Agresti (2000) were unable to draw the conclusion that increased model complexity implies an increase in predictive correlation. Since, as mentioned in Section 1, increased model complexity cannot hurt the best predictor, the squared predictive correlation cannot decrease under increased model complexity.

3. General Comments on Estimation

The usual definition of R^2 from linear models provides an estimate of the squared multiple correlation coefficient, which is a fundamental measure of how well the best linear predictor works. A strictly analogous fundamental measure of how well the best predictor works is the maximum squared predictive correlation. For nonlinear (as well as linear) models, it is therefore appropriate to define R^2 as an estimate of the squared predictive correlation.

The general regression model is $\mathrm{E}(y|x) = m(x)$. Suppose we have a random sample (x_i', y_i), $i = 1, \ldots, n$. A generalized linear model assumes a distribution for y given x and that $\mathrm{E}(y_i|x_i) = m(\alpha + x_i'\beta)$ for known m and unknown α and β. Here m is just the inverse of the link function. The standard nonparametric regression model is $y_i = m(x_i) + \varepsilon_i$ where, conditional on the x_is, the ε_is are independent with mean 0 and variance σ^2. In nonparametric regression, m is unknown. The standard nonlinear regression model uses $m(x_i) = m(x_i; \alpha, \beta)$ where m is known but α and β are unknown. In linear regression, $m(x_i) = \alpha + x_i'\beta$, again with α and β unknown. The conditional mean structure of all three parametric models is that of the nonlinear regression model: $m(x) = m(x; \alpha, \beta)$, m known.

Best prediction theory treats $m(x)$ as a known function, so for models involving α and β it treats them as known. With m known, an obvious estimate of the highest squared predictive correlation is the squared sample correlation of the pairs $(y_i, m(x_i))$. In practice, $m(x)$ must be estimated with $\hat{m}(x)$. This is either done nonparametrically, or by estimating α and β and substituting them into the known function $m(x; \alpha, \beta)$. The highest squared predictive correlation is estimated by the squared sample correlation of the pairs $(y_i, \hat{m}(x_i))$.

In practice, we do not know the conditional expectation $m(x)$. We simply create some model for the conditional expectation, say, $f(x)$. The model can be nonparametric, nonlinear, generalized linear, or linear. Again, if f is known, the pairs $(y_i, f(x_i))$ provide an estimate of the squared predictive correlation. If f has to be estimated, we define $\tilde{y}_i \equiv \hat{f}(x_i)$ and use the pairs $(y_i, \tilde{y}_i)$. This is estimating the squared predictive correlation of $f(x)$, which we know is no greater than the squared predictive correlation provided by the best predictor

$m(x)$. Zheng and Agresti (2000) investigated the bias that results from using the same data to estimate both the predictor and predictive correlation. In addition to the estimate discussed here, they considered jack-knife, modified jack-knife, and cross-validation estimates and found the naive estimate to be quite good.

As we have seen, the squared predictive correlation does not actually measure how well $\hat{f}(x)$ predicts y, it measures only the potential of $\hat{f}(x)$ to predict y. To ensure that we have a good predictor based on $\hat{f}(x)$, we need to regress the y_i's on the $\hat{f}(x_i)$'s to get a new predictor $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \tilde{y}_i$. Since $\hat{y}_i$ is based on the theory of best linear prediction, the appropriate estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ would seem to be least squares estimates, see Christensen (2002, Sec. 6.3), regardless of whether the conditional distribution of y given x has constant variance. (If f is known, least squares are certainly appropriate.) With y given x being homoscedastic normal data, standard tests of $\beta_0 = 0$ and $\beta_1 = 1$ might be used to determine whether this linearization is really necessary.

Based on the theory of Section 2, there are three equivalent ways to define the squared predictive correlation,

$$\frac{\sigma^2_{y\tilde{y}}}{\sigma_{yy}\sigma_{\tilde{y}\tilde{y}}} = \frac{\sigma^2_{y\hat{y}}}{\sigma_{yy}\sigma_{\hat{y}\hat{y}}} = \frac{\sigma_{\hat{y}\hat{y}}}{\sigma_{yy}}.$$

Using least squares to create the linearized predictor $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \tilde{y}_i$, there are four equivalent ways to estimate the squared predictive correlation,

$$R^2 = \frac{s^2_{y\tilde{y}}}{s_{yy}s_{\tilde{y}\tilde{y}}} = \frac{s^2_{y\hat{y}}}{s_{yy}s_{\hat{y}\hat{y}}} = \frac{s_{\hat{y}\hat{y}}}{s_{yy}}$$

as well as $R^2 = 1 - [s_{\hat{\varepsilon}\hat{\varepsilon}}/s_{yy}]$ where the $\hat{\varepsilon}_i$s are the residuals from regressing the y_is on the $\tilde{y}_i$s. The last two of these forms are simply $R^2 = SSReg/SSTot$ and $R^2 = 1 - [SSE/SSTot]$ from the regression on the $\tilde{y}_i$s.

The squared predictive correlation is an overall measure of predictive ability. It is based on the joint distribution of x and y, not the conditional distribution. As such, the x_is must be considered a random sample from some population. The nature of that population should be given careful consideration. For example, in logistic regression, the size of the estimated squared predictive correlation will depend crucially on where the x_is are sampled, cf. also Zheng and Agresti (2000).

The quality of predictions often change depending on x. In logistic regression, x values corresponding to true probabilities near 0 or 1 predict y very accurately, but x values corresponding to true probabilities near .5 can never predict y with high accuracy. R^2 is a measure of overall predictive ability. If the x_i are sampled from areas corresponding to probabilities near .5, R^2 will be relatively small. If the sample of x_is contains mostly vectors corresponding to probabilities near 0 or 1, R^2 will be large. Note that in logisitic and other forms of binomial regression, it is almost impossible to achieve perfect prediction, so even with a perfect model the squared predictive correlation will be less than 1.

It is often suggested that comparing the R^2 of (even linear) models from data that have different x_i values is inappropriate. In the present context in which the x_is are considered

a random sample from some population, very different collections of x_i values suggest that the two sets of x_is are being sampled from different populations, in which case it is clearly inappropriate to compare R^2 values.

The assumption that the (x_i', y_i)s are iid is important. Designed experiments might be thought of as samples from a discrete distribution on x. Correlated data is more difficult to treat.

4. Regression Through the Origin

A subject that has long been of some controversy has been how to properly define R^2 for regression through the origin. We will see that the issue is not how to define the squared multiple correlation coefficient, but rather how to estimate it. In regression through the origin, we *assume* the $m(x) = x'\beta$ for some vector β. With the techniques used in proving Theorem 2, it is not difficult to see that

$$\mathrm{E}[y - x'\beta]^2 = \mathrm{E}[y - x'\mu_{xx}^{-1}\mu_{xy}]^2 + \mathrm{E}[x'\mu_{xx}^{-1}\mu_{xy} - x'\beta]^2, \tag{4.1}$$

where $\mu_{xy} \equiv \mathrm{E}[xy]$ and $\mu_{xx} \equiv \mathrm{E}[xx']$. From equation (4.1) it is clear that the best predictor in this class and therefore, by assumption, the best predictor is

$$m(x) = x'\mu_{xx}^{-1}\mu_{xy}. \tag{4.2}$$

The key fact is that taking expectations on both sides of (4.2) gives

$$\mu_y = \mu_x'\mu_{xx}^{-1}\mu_{xy}. \tag{4.3}$$

We need to use this fact in estimating the squared multiple correlation.

Using equation (4.3) and the general theory, we find that

$$\mathrm{Var}[x'\mu_{xx}^{-1}\mu_{xy}] = \mathrm{Cov}[y, x'\mu_{xx}^{-1}\mu_{xy}] = \mu_{yx}\mu_{xx}^{-1}\mu_{xy} - \mu_y^2,$$

so the squared multiple correlation is

$$\frac{\mu_{yx}\mu_{xx}^{-1}\mu_{xy} - \mu_y^2}{\sigma_{yy}} = \frac{\mu_{yx}\mu_{xx}^{-1}\mu_{xy} - \mu_y^2}{\mu_{yy} - \mu_y^2}.$$

To estimate this quantity, we set up the usual linear model $Y = X\beta + e$, let $M = X(X'X)^{-1}X'$ denote the perpendicular projection operator (ppo) onto the column space of X, $C(X)$. Like all ppos, M is idempotent and symmetric, i.e., $MM = M$ and $M = M'$. The vector of 1's is denoted J with $\frac{1}{n}JJ' = J(J'J)^{-1}J'$ the ppo onto $C(J)$. Define estimates

$$\hat{\mu}_{yy} \equiv \frac{1}{n}Y'Y \quad \hat{\mu}_{xy} \equiv \frac{1}{n}X'Y \quad \hat{\mu}_{xx} \equiv \frac{1}{n}X'X.$$

The trick is to use estimates of μ_y and μ_x that satisfy (4.3) when μ_{yx} and μ_{xx} have been replaced by their estimates. The sample means $\bar{y}_\cdot$ and $\bar{x}_\cdot$ do not meet this criterion. We propose to use

$$\hat{\mu}_y \equiv \frac{1}{n} Y'MJ$$

and

$$\hat{\mu}_x \equiv \frac{1}{n} X'MJ = \frac{1}{n} X'J = \bar{x}_\cdot.$$

Clearly, $\mathrm{E}(\hat{\mu}_x) = \mu_x$ but note that from (4.2) and (4.3)

$$\mathrm{E}(\hat{\mu}_y) = \mathrm{E}_x \mathrm{E}_{y|x} \left[\frac{1}{n} Y'MJ \right] = \mathrm{E}_x \left[\frac{1}{n} \mu_{yx} \mu_{xx}^{-1} X'MJ \right] = \mathrm{E}_x \left[\mu_{yx} \mu_{xx}^{-1} \bar{x}_\cdot \right] = \mu_{yx} \mu_{xx}^{-1} \mu_x = \mu_y.$$

With these estimates

$$\begin{aligned}
R^2 &= \frac{\hat{\mu}_{yx} \hat{\mu}_{xx}^{-1} \hat{\mu}_{xy} - \hat{\mu}_y^2}{\hat{\mu}_{yy} - \hat{\mu}_y^2} \\
&= \frac{Y'X(X'X)^{-1}X'Y - \frac{1}{n} Y'MJJ'MY}{Y'Y - \frac{1}{n} Y'MJJ'MY} \\
&= \frac{Y'[M - M(\frac{1}{n} JJ')M]Y}{Y'[I - M(\frac{1}{n} JJ')M]Y} \\
&= \frac{Y'M[I - (\frac{1}{n} JJ')]MY}{Y'(I-M)Y + Y'M[I - (\frac{1}{n} JJ')]MY} \\
&= \frac{s_{\hat{y}\hat{y}}}{SSE/(n-1) + s_{\hat{y}\hat{y}}}
\end{aligned}$$

Obviously, $R^2 = 1$ when there is no error in the predictions and $R^2 = 0$ when the predictions do not change with the x_is.

The literature is full of ideas for creating R^2 like measures for problems other than linear models with an intercept. In the case of regression through the origin, one such measure is to use the usual definition of $SSReg/SSTot$ except not correct either sum of squares for the mean, i.e, use $\tilde{R}^2 = Y'MY/Y'Y$. This can be viewed as an estimate of $\tilde{\rho}^2 = \mu_{yx} \mu_{xx}^{-1} \mu_{xy} / \mu_{yy}$ which is *not* the squared predictive correlation. Both $\tilde{R}^2$ and $\tilde{\rho}^2$ will be 1 when the regression through the origin gives perfect prediction and 0 when the model has no predictive ability. Note that when there is no predictive ability, $m(x) = \mu_y$ and since, by assumption, $m(0) = 0$, $m(x) = 0$. On the other hand, there is nothing to keep one from using $\tilde{R}^2$ and $\tilde{\rho}^2$ for linear models with an intercept, in which case $\tilde{R}^2$ and $\tilde{\rho}^2$ are 1 for perfect prediction but would not be 0 whenever there is no predictive ability. In particular, $\tilde{\rho}^2$ will only be 0 if $m(x) = 0$.

While $\tilde{\rho}^2$ may be a reasonable measure for comparing alternative models of regression through the origin, the argument above illustrates that it does not extend to comparing regression through the origin models to other models, either linear regression not through the origin or other nonlinear models. Theoretically, squared predictive correlation has no such drawbacks because regression through the origin is merely a special case of best prediction. The only difficulty with regression through the origin is in finding an appropriate estimate of the squared predictive correlation.

5. Method of Moments Estimation

General prediction theory is based solely on the existence of first (conditional) and second moments. The only basis it provides for parameter estimation is method of moments (MOM) estimation. A natural way to perform estimation in general prediction theory is through linear estimating equations, cf. McCullagh and Nelder (1989, Section 9.5). Consider a parametric model $\mathrm{E}(y|x) \equiv m(x;\beta)$ where m is known but β is an unknown r vector.

We assume that β is both well defined, i.e., that $m(x;\beta) = m(x;\beta_*)$ for all x implies that $\beta = \beta_*$ and well defined in the estimation problem, i.e., that with $m(X;\beta)$ denoting an n vector that applies $m(\cdot;\beta)$ to each row of X, $m(X;\beta) = m(X;\beta_*)$ implies that $\beta = \beta_*$. Obviously, if β is well defined in the estimation problem, β is well defined. (These properties for all x need only apply almost surely.)

If β is well defined in the estimation problem and $\mathrm{E}(Y|X)$ is known, $\mathrm{E}(Y|X) - m(X;\beta) = 0$, so we can identify β by solving

$$H'[\mathrm{E}(Y|X) - m(X;\beta)] = 0$$

for any $n \times r$ matrix H as long as the r equations are not redundant. H can be as simple as $H' = [I_r, 0]$ but H may also depend on X and β. H cannot depend on Y. Any such solution allows us to find β exactly, so how we pick H is irrelevant.

Unfortunately, we don't know $\mathrm{E}(Y|X)$, but we can estimate β using the linear estimating equation

$$H'[Y - m(X;\beta)] = 0.$$

Now the challenge is to find a matrix H with good statistical properties. It is well known that, conditional on X, solutions to the quasi-likelihood equations are asymptotically optimal among all solutions to linear estimating equations. The quasi-likelihood equations are

$$d_\beta m(X;\beta)' V^{-1}(X;\beta)[Y - m(X;\beta)] = 0,$$

where $d_\beta m(X;\beta)$ is an $n \times r$ matrix with ith row $d_\beta m(x_i;\beta)$ which is the row vector of partial derivatives of m with respect to β evaluated at x_i and $V^{-1}(X;\beta)$ is a diagonal matrix having elements $V(x_i;\beta) \equiv \mathrm{Var}(y|x_i)$. Any additional parameters that relate solely to the variance must be estimated separately or iteratively.

The prediction theory used here is based only on first and second moments and conditional moments. Typically, to get better estimates, one needs to make stronger assumptions. Both generalized linear models and nonlinear regression make explicit assumptions about the distribution of y given x, thus allowing one to find maximum likelihood estimates, but the maximum likelihood estimates agree with the solutions to the corresponding quasi-likelihood equations.

These estimation methods have implications for the idea of linearizing predictors. The best predictor is unchanged by linearizing it, but in practice, the primary way to see whether linearization will improve prediction is to try it and see. Since these estimation methods are directed at finding $\hat{\beta}$ to be consistent with $m(x;\hat{\beta})$ being the best predictor, there should be

little room for improvement by linearizing the fitted regression function $m(x;\hat{\beta})$. In the case of $\hat{m}$ obtained from linear regression, linearizing makes no change in the predictors, no matter how bad the model is, that is, no matter how far $\mathrm{E}(y|x)$ is from the chosen form $m(x;\beta)$. For generalized linear models and nonlinear regression models, it would probably take an extraordinarily bad choice for m before linearizing would show that there is a problem. However, for complex computer models, in which the estimation method is not so clear, linearizing may be valuable. In particular, linearization will have no effect at all whenever $[J,\hat{m}]'(Y-\hat{m})=0$, because then $\bar{y}_\cdot = \sum_{i=1}^n \hat{m}_i/n$ and $s_{\hat{m}y} = s_{\hat{m}\hat{m}}$, so the regression coefficient for linearization will be 1 and the intercept will be 0. From Proposition 3, $\mathrm{Cov}[m(x), y-m(x)]$ should be 0, and here $\hat{m}'(Y-\hat{m})/n$ is just an estimate of that covariance. So the effect of linearization in nonlinear models will depend on how close $[J,\hat{m}]'(Y-\hat{m})$ is to 0.

EXAMPLE: For the O-ring data, regressing the failures y on the predicted values $\tilde{y}$ from the logistic regression on temperature gives $\hat{y} = -0.0073 + 1.0239\tilde{y}$, so clearly there is little to be gained by linearizing the maximum likelihood fit. Similar results occur from regressing the failures on the predicted values from the quadratic model and both probit models.

Recently, Nayak (2002) discussed a Cramer-Rao lower bound for prediction problems and methods of finding best unbiased predictors by attaining the bound.

6. Error Estimation and Residuals

With m known, it is a simple matter to estimate the prediction error variance $\mathrm{Var}_{xy}[y-m(x)]$: use $\sum_{i=1}^n [y_i - m(x_i)]^2/n$. When estimating m using, say, r parameters, a reasonable estimate is $\sum_{i=1}^n [y_i - \hat{m}(x_i)]^2/(n-r)$. Note that if the conditional variance of y given x is a constant σ^2 that does not depend on x, $\sum_{i=1}^n [y_i - \hat{m}(x_i)]^2/(n-r)$ provides an estimate of σ^2 because

$$\mathrm{Var}_{xy}[y-m(x)] = \mathrm{E}_{xy}[y-m(x)]^2 = \mathrm{E}_x \mathrm{E}_{y|x}[y-m(x)]^2 = \mathrm{E}_x \sigma^2 = \sigma^2.$$

Constant conditional variance is a standard assumption for linear and nonlinear regression, but is not standard in many generalized linear models such as logistic regression. On the other hand, heteroscedasticity in the conditional distribution plays no role in the estimation because we are estimating a property of the joint distribution, not a property of the conditional distribution. The presumption is that the pairs (x_i, y_i) are iid, but clearly that does not imply that the $y_i|x_i$ are iid or even homoscedastic.

More generally, if $\mathrm{Var}_{y|x}(y) = \sigma^2 w(x)$ for some fixed function $w(\cdot)$, an estimate of σ^2 is $\sum_{i=1}^n [y_i - \hat{m}(x_i)]^2/w(x_i)(n-r)$. This works if $w(\cdot)$ is a known function or, if $w(\cdot)$ is a known function of $m(\cdot)$, $w(\cdot)$ can be estimated. Note that having heterogenous conditional variances does not affect the general prediction theory, even though they would affect the process of estimating the parameters in a model for the best predictor. If $w(x)$ is unknown, $\sigma^2 w(x)$ is ill-defined unless one value of $w(x)$ is taken as known. With this proviso, $\sigma^2 w(x)$ is the best predictor of the random variable $[y-m(x)]^2$, so all of the standard methods of regression can be used to estimate $\sigma^2 w(x)$.

Although we have discussed R^2 as a tool in model fitting, the prediction theory fundamentally evaluates a predictor $\tilde{y}(x)$ based on $\mathrm{E}_{xy}[y-\tilde{y}(x)]^2$. One could argue that we should base model evaluations on estimates of this quantity. The obvious estimate is $\sum_{i=1}^{n}[y_i-\tilde{y}(x_i)]^2/n$, but as discussed earlier, this will be overly optimistic when using the data to estimate $\tilde{y}$. As suggested earlier, one possible way to adjust for fitting different numbers of predictor variables is to use $\sum_{i=1}^{n}[y_i-\tilde{y}(x_i)]^2/(n-r)$.

EXAMPLE: The *Challenger* data involves 23 flights of which 7 experienced O-ring failures. The mean squared prediction errors for the logit and probit models using temperature alone are .138527 and .140069, respectively. Using quadratic models the squared prediction errors are .129888 and .130973, respectively. Qualitatively, these tell the same story as the R^2 values. The corresponding value for the constant predictor $\hat{p} = 7/23$ is .211720. If we adjust the denominator for the number of predictor variables, the numbers become .151439, .153409, .149371, .150615, .221344, respectively.

Various other ideas have been proposed to deal with the bias. Breiman (2000) and others have suggested estimating $\tilde{y}$ using a randomly selected, say, 90% of the data, and computing $\sum_{i=1}^{n}[y_i-\tilde{y}(x_i)]^2/n$ on the other 10%. Moreover, Breiman suggests doing the random selection many times and averaging the results. Another idea is to simply view $\theta \equiv \mathrm{E}_{xy}[y-\tilde{y}(x)]^2$ as a parameter, $\hat{\theta} \equiv \sum_{i=1}^{n}[y_i-\tilde{y}(x_i)]^2/n$ as an estimate, and jackknife the estimate to reduce bias. It is not difficult to see that for linear models, the jackknifed estimator of θ is the mean squared error times the sum of squares of the standardized residuals divided by the sample size, see Christensen (2002, Sec. 13.5) for background. Note that this is different from the PRESS statistic or the average of the PRESS statistic.

EXAMPLE: Atkinson (1985) and Hader and Grandage (1958) presented Prater's data on gasoline. The variables are y, the percentage of gasoline obtained from crude oil; x_1, the crude oil gravity °API; x_2, crude oil vapor pressure measured in lbs/in^2; x_3, the temperature, in °F, at which 10% of the crude oil is vaporized; and x_4, the temperature, in °F, at which all of the crude oil is vaporized. Table 2 gives the four subset models with the lowest C_p statistics along with their R^2, adjusted R^2, MSE, and jackknifed prediction error values (JK).

7. Residual Plots

Figure 1 contained a residual plot for the Hooker data of regression y (pressure) on x (temperature). It shows a nonrandom pattern. We now provide a theoretical justification for looking at residual plots and for using them as a tool for adding linear predictors.

Theorem 6. Suppose $\tilde{y}(x)$ is any predictor with $\mathrm{E}[\tilde{y}(x)] = \mu_y$, then $\mathrm{Cov}[f(x), y-\tilde{y}(x)] = 0$ for any function f if and only if $m(x) = \tilde{y}(x)$.

From Theorem 6, the theoretical residuals $y - m(x)$ are uncorrelated with any function $f(x)$. Thus, for any function f, plot the pairs $[f(x_i), y_i - \tilde{y}(x_i)]$ to see if it gives a nice

structureless plot. If there is a correlation, obviously $\tilde{y}$ is not acting like the best predictor. We should be able to improve the predictor by a multiple linear regression of the y_is on $\tilde{y}(x_i)$s and any $f(x_i)$s that are correlated with the residuals.

Conversely, again by Theorem 6, if we have the wrong mean function, that is, if $\tilde{y}(x) \neq m(x)$, there exist functions $f(x)$ that have a nonzero correlation with the residuals $y - \tilde{y}(x)$ and we can hope to find an $f(x)$ using a residual plot.

EXAMPLE: Although Figure 1 displays structure, the theory of linear models establishes that the sample correlation between the two variables plotted is 0. Figure 4 contains a plot of the residuals versus $f(x) = (x - \bar{x})^2$. It shows a clear linear relationship, suggesting that we may improve prediction by adding $f(x)$ as a linear predictor, that is, by fitting a quadratic model.

8. Conclusions and Discussion

Measuring predictive ability by the squared correlation between observations and predictions is a simple and obviously good idea. At a December, 1999, National Academy of Sciences workshop in Santa Fe, NM on Statistical Approaches for the Evaluation of Complex Computer Models, numerous nonstatistical scientists justified the quality of their complex prediction models by plotting y versus $\tilde{y}$, see Berk et al. (2002). We simply propose summarizing this plot by computing the squared correlation. Outside of Mittlbock and Schemper (1996) and Zheng and Agresti (2000), this idea has rarely appeared in the statistics literature. In the collective works Agresti (1986, 1990, 1996), Bates and Watts (1988), Cameron and Windmeijere (1997), Christensen (1997), Draper and Smith (1981), Eubank (1988), Green and Silverman (1994), Hart (1997), Hosmer and Lemeshow (1989), McCullagh and Nelder (1989), Menard (2000), and Seber and Wild (1989), for all of whom the idea of extending R^2 to nonlinear models is relevant, the only discussion (that I could find) is Draper and Smith *mentioning* that this idea gives R^2 for linear models and Menard (2000) dismissing it because Kvalseth (1985) dismissed it.

Kvalseth (1985) argued, "Such a correlation interpretation of R^2 would not seem to be particularly attractive or useful on intuitive grounds as a goodness-of-fit measure for nonlinear models. Furthermore, the use of R^2 for nonlinear models can produce potentially misleading results since it is clearly possible for y and $[\tilde{y}]$ to be highly correlated even if their corresponding values deviate substantially." While both of these criticisms are valid, both are easily dealt with. First, the squared correlation between y and $\tilde{y}$ is a very intuitive measure of the predictive ability of a model but as mentioned earlier, it is not a direct measure of goodness of fit. R^2 does play a key role in assessing *relative* goodness of fit. Second, while it is true that a high R^2 can occur with very bad predictors $\tilde{y}$, this is easily remedied by simply obtaining new "linearized" predictors $\hat{y}$ from regressing y on $\tilde{y}$. Moreover, there are some applications, for example in psychometrics and data mining, that do not require such linearization because any linear transformation of the predictor is as good as any other. For such problems, R^2 is appropriate but squared error loss is not.

Other R^2 measures share with squared predictive correlation some ability to compare nonnested models. For example, if one identifies a saturated model M_s, a model M_0 that characterizes no predictive ability, and a model of interest M, a commonly used measure based on the likelihood function $L(\cdot)$ is

$$R_L^2 = \frac{L(M) - L(M_0)}{L(M_s) - L(M_0)}.$$

The saturated model gives "perfect" predictions (for the data used to estimate the parameters), so R_L^2 will be 0 for no predictive ability and 1 for perfect prediction. Of course the saturated model does not really give perfect prediction, that is only an artifact of using the same data both to estimate parameters and to evaluate predictive ability. In fact, the best predictor will only give perfect prediction in the degenerate case where $y = \mathrm{E}(y|x)$ a.s. The very use of the likelihood function makes R_L^2 a measure of how well the data estimate parameters in various models rather than a true measure of predictive ability. It is not clear what fundamental population parameter R_L^2 might estimate. Squared predictive correlation also applies in this situation and has a clearer theoretical basis, one that is not tied to the specific method of parameter estimation (which is maximum likelihood). See also Mittlbock and Schemper (1996) and Zheng and Agresti (2000).

Additionally, when comparing different transformations, say y and $\log(y)$, the models typically suggest different distributions for $y|x$, so, unlike squared predictive correlation, likelihood based comparisons across models are inappropriate. In reality, there is only one conditional distribution for $y|x$, the assumed models are changing, not the distribution. Regardless of the correct conditional distribution, squared predictive correlation gives a useful measure of predictive ability and relative goodness of fit.

The results in Section 2 are based on squared error prediction loss. It might be possible to develop an alternative general prediction theory based on another loss function such as entropy, but squared error is the standard loss function. For particular applications, it may be appropriate to develop measures of predictive ability using a predictive loss function that depends on the conditional distribution of y given x, however, by their very nature, such techniques will not apply beyond the specific application and not contribute to general prediction theory. It should also be noted that in general prediction theory, predictors $f(\cdot)$ are treated as known and the issue is to find the best (linear) predictor. In this context it makes no sense to let the predictive loss function depend on the estimation method that one might choose to employ for approximating the optimal predictor.

We have seen that generalizing the linear prediction concept of squared multiple correlation to the general prediction concept of squared predictive correlation provides a theoretical basis for estimating the predictive ability of a nonlinear model by looking at the squared correlation between the observations and their model based predictions. This extends the theoretical basis behind the standard measure R^2 from linear models to nonlinear models.

The absolute size of the squared predictive correlation for $\hat{y}$ has little to do with how well the model fits. Estimated squared predictive correlations can be small for perfect models and large for demonstrably incorrect models. A prediction model $f(x)$ is a good fit if $f(x)$

closely approximates the best predictor $m(x)$ (see Theorem 1). The prediction model can be improved by linearizing $f(x)$. How well the linearized model fits is a function of how close the squared predictive correlation of $f(x)$ is to the squared predictive correlation of $m(x)$. This is something we cannot know because we do not know m, we can only model it. However, the absolute size of the squared predictive correlation of $f(x)$ does tell us about how well the linearized model will predict.

With linearized predictors the relative size of the squared predictive correlation provides information on which of two models is closer to the ideal. In linear regression, for a fixed number of predictors, the standard model selection criteria C_p and adjusted R^2 both give their highest rankings to the models with the largest R^2. In linear models it is standard (and good) advice that with different numbers of predictors, one should not simply look at R^2 because R^2 can only increase or stay the same when new variables are added. This is a problem with using R^2 as an *estimate* of the squared predictive correlation. Theoretically, adding another predictor variable can never hurt prediction. The worst thing that can happen is that the theoretical regression coefficient for the new variable may be 0, in which case no harm has been done. The harm occurs in estimating regression coefficients that are close to 0. More generally, with linearized predictors, finding a predictor that increases the theoretical squared predictive correlation only helps prediction. But when using an estimated predictor, and an estimated squared predictive correlation, adding parameters that need to be estimated when the result is only a small increase in estimated squared predictive correlation, can be counterproductive. The issue of how C_p and adjusted R^2 penalize models for including additional variables is also really a question of how best to estimate the ideal function $m(x)$.

Finally, it should be noted that this approach requires y to be a random variable, a requirement that may preclude its use with some multinomial response models.

Appendix: Proofs

First, we show that for a linear model the sample correlation between y and the predicted values $\hat{y}$ equals the sum of squares regression divided by the sum of squares total. That is followed by proofs of propositions and theorems.

Consider a linear model $Y = X\beta + e$ in which the first column of X is the vector of 1's, J. M is the ppo onto $C(X)$, $\frac{1}{n}JJ'$ is the ppo onto $C(J)$, and the vector of predicted values is $\hat{Y} = MY$. A vector that consists entirely of the mean of the y_is is $\bar{y}_\cdot J = \frac{1}{n}JJ'Y$. The sample variance of the y_is is

$$s_y^2 \equiv s_{yy} = \frac{1}{n-1}SSTot = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_\cdot)^2$$
$$= \frac{1}{n-1}[(I - \frac{1}{n}JJ')Y]'[(I - \frac{1}{n}JJ')Y] = \frac{1}{n-1}Y'(I - \frac{1}{n}JJ')Y.$$

The sample variance of the $\hat{y}_i$s is

$$s^2_{\hat{y}} \equiv s_{\hat{y}\hat{y}} = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_\cdot)^2 = \frac{1}{n-1}[(M - \frac{1}{n}JJ')Y]'[(M - \frac{1}{n}JJ')Y]$$
$$= \frac{1}{n-1}Y'(M - \frac{1}{n}JJ')Y = \frac{1}{n-1}SSReg.$$

Now, using the fact that $J \in C(X)$, so $MJ = J$, the sample covariance between the y_is and the $\hat{y}_i$s is

$$\begin{aligned} s_{y\hat{y}} &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_\cdot)(\hat{y}_i - \bar{y}_\cdot) \\ &= \frac{1}{n-1}[(I - \frac{1}{n}JJ')Y]'[(M - \frac{1}{n}JJ')Y] \\ &= \frac{1}{n-1}Y'[M - \frac{1}{n}JJ'M - \frac{1}{n}JJ' + \frac{1}{n}JJ'\frac{1}{n}JJ']Y \\ &= \frac{1}{n-1}Y'(M - \frac{1}{n}JJ')Y \\ &= \frac{1}{n-1}SSReg. \end{aligned}$$

It follows that

$$R^2 \equiv \frac{s^2_{y\hat{y}}}{s^2_y s^2_{\hat{y}}} = \frac{[SSReg/(n-1)]^2}{[SSTot/(n-1)][SSReg/(n-1)]} = \frac{SSReg}{SSTot}.$$

Proof of Proposition 3. Recall that from the definition of conditional expectation $\mathrm{E}[m(x)] = \mu_y$.

$$\begin{aligned} \mathrm{Cov}[y, \tilde{y}(x)] &= \mathrm{E}_{yx}[(y - \mu_y)\tilde{y}(x)] \\ &= \mathrm{E}_x\mathrm{E}_{y|x}[(y - m(x) + m(x) - \mu_y)\tilde{y}(x)] \\ &= \mathrm{E}_x[(m(x) - \mu_y)\tilde{y}(x)] \\ &= \mathrm{Cov}[m(x), \tilde{y}(x)]. \end{aligned}$$

Proof of Theorem 4. By Cauchy-Schwartz, $(\sigma_{m\tilde{y}})^2 \leq \sigma_{mm}\sigma_{\tilde{y}\tilde{y}}$, so $(\sigma_{m\tilde{y}})^2/\sigma_{\tilde{y}\tilde{y}} \leq \sigma_{mm}$. Using Proposition 3

$$\frac{(\sigma_{y\tilde{y}})^2}{\sigma_{yy}\sigma_{\tilde{y}\tilde{y}}} = \frac{(\sigma_{m\tilde{y}})^2}{\sigma_{yy}\sigma_{\tilde{y}\tilde{y}}} \leq \frac{\sigma_{mm}}{\sigma_{yy}}.$$

The result follows from the last part of Proposition 3.

Proof of Theorem 5. $\sigma_{\hat{y}_1\hat{y}_1}/\sigma_{yy} < \sigma_{\hat{y}_2\hat{y}_2}/\sigma_{yy}$ if and only if $\sigma_{\hat{y}_1\hat{y}_1} < \sigma_{\hat{y}_2\hat{y}_2}$ if and only if $\sigma_{yy} - \sigma_{\hat{y}_2\hat{y}_2} < \sigma_{yy} - \sigma_{\hat{y}_1\hat{y}_1}$.

PROOF OF THEOREM 6. If $m(x) = \tilde{y}(x)$, the fact that $\text{Cov}[f(x), y - m(x)] = 0$ for any function f is an immediate consequence of Proposition 3.

Now suppose that $\text{E}[\tilde{y}(x)] = \mu_y$ and $\text{Cov}[f(x), y - \tilde{y}(x)] = 0$ for any function f. In particular, using Proposition 3,

$$0 = \text{Cov}[\tilde{y}(x), y - \tilde{y}(x)] = \sigma_{\tilde{y}y} - \sigma_{\tilde{y}\tilde{y}} = \sigma_{\tilde{y}m} - \sigma_{\tilde{y}\tilde{y}}$$

and

$$0 = \text{Cov}[m(x), y - \tilde{y}(x)] = \sigma_{my} - \sigma_{m\tilde{y}} = \sigma_{mm} - \sigma_{m\tilde{y}}.$$

Therefore,

$$\sigma_{mm} = \sigma_{m\tilde{y}} = \sigma_{\tilde{y}\tilde{y}}.$$

This implies that the correlation between $m(x)$ and $\tilde{y}(x)$ is 1, so they are linear functions of each other. Moreover, because they have the same mean and variance, the functions must be identical (almost surely).

ACKNOWLEDGEMENT

Thanks to Peter Westfall for valuable comments.

REFERENCES

Agresti, Alan (1986). "Applying R^2-type Measures to Ordered Categorical Data," *Technometrics*, **28**, 133-138.

Agresti, Alan (1990). *Categorical Data Analysis*. New York: John Wiley and Sons.

Agresti, Alan (1996). *An Introduction to Categorical Data Analysis*. New York: John Wiley and Sons.

Atkinson, A. C. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford University Press, Oxford.

Bates, Douglas M. and Watts, Donald G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley and Sons, New York.

Berk, R. A., Bickel, P., Cambell, K., Fovell, R., Kelly-McNulty, S., Kelly, E., Linn, R., Park, B., Perelson, A., Rouphail, N., Sacks, J., and Schoenberg, F. "Workshop on Statistical Approaches for the Evaluation of Complex Computer Models." *Statistical Science*, **17**, 173-192.

Blattenberger, G. and Lad, F. (1985). "Separating the Brier score into calibration and refinement components: A graphical exposition," *The American Statistician*, **39**, 26-32.

Breiman, Leo (2000). "Statistical Modeling: The Two Cultures," with discussion. *Statistical Science*, **16**, 199-231.

Cameron, A. Colin and Windmeijer, A. G. (1997). An R^2 measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, **77**, 329- 342.

Christensen, Ronald (1996). *Analysis of Variance, Design, and Regression: Applied Statistical Methods*. Chapman and Hall, London.

Christensen, Ronald (1997). *Log-Linear Models and Logistic Regression*, Second Edition. Springer-Verlag, New York.

Christensen, Ronald (2002). *Plane Answers to Complex Questions: The Theory of Linear Models*, Third Edition. Springer-Verlag, New York.

Draper, N. and Smith, H. (1981). *Applied Regression Analysis*, Second Edition. John Wiley and Sons, New York.

Eubank, Randall L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London.

Hader, R. J. and Grandage, A. H. E. (1958). "Simple and Multiple Regression Analyses," in *Experimental Designs in Industry*, edited by V. Chew, pp. 108–137. John Wiley and Sons, New York.

Hart, Jeffrey D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer-Verlag, New York.

Hosmer, David W. and Lemeshow, Stanley (1989). *Applied Logistic Regression*. New York: John Wiley and Sons.

Kvalseth, T. O. (1985). "Cautionary Note about R^2," *The American Statistician*, **39**, 279-285.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Second Edition. Chapman and Hall, London.

Menard, Scott. (2000). "Coefficients of Determination for Multiple Logistic Regression Analysis," *The American Statistician*, **54**, 17-24.

Mittlbock, M. and Schemper, M. (1996). "Explained variation for logistic regression," *Statistics in Medicine*, **15**, 1987-1997.

Nayak, Tapan K. (2002). "Rao-Cramer Type Inequalities for Mean Squared Error of Prediction," *The American Statistician*, **56**, 102-106.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Second Edition. John Wiley and Sons, New York.

Schmid, C. H. and Griffith, J. L. (1998). "Multivariate Classification Rules: Calibration and Discrimination," in *Encyclopedia of Biostatistics*, edited by P. Armitage and T. Colton, **4**, 2844-2850. John Wiley and Sons, Chichester.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression*. John Wiley and Sons, New York.

Weisberg, S. (1985). *Applied Linear Regression*, Second Edition. John Wiley and Sons, New York

Zheng, B, and Agresti, A. (2000). "Summarizing the predictive power of a generalized linear model," *Statistics in Medicine*, **19**, 1771-1781.

Case	Temperature	Pressure	Case	Temperature	Pressure
1	180.6	15.376	17	191.1	19.490
2	181.0	15.919	18	191.4	19.758
3	181.9	16.106	19	193.4	20.480
4	181.9	15.928	20	193.6	20.212
5	182.4	16.235	21	195.6	21.605
6	183.2	16.385	22	196.3	21.654
7	184.1	16.959	23	196.4	21.928
8	184.1	16.817	24	197.0	21.892
9	184.6	16.881	25	199.5	23.030
10	185.6	17.062	26	200.1	23.369
11	185.7	17.267	27	200.6	23.726
12	186.0	17.221	28	202.5	24.697
13	188.5	18.507	29	208.4	27.972
14	188.8	18.356	30	210.2	28.559
15	189.5	18.869	31	210.8	29.211
16	190.6	19.386			

Table 1: Hooker data.

		Adj.				Included variables			
Vars.	R^2	R^2	C_p	$\sqrt{MSE}$	JK	x_1	x_2	x_3	x_4
2	95.2	94.9	8.2	5.88	5.88			X	X
3	95.9	95.5	5.2	5.21	5.08	X		X	X
3	95.5	95.0	8.2	5.74	5.65		X	X	X
4	96.2	95.7	5.0	4.99	4.78	X	X	X	X

Table 2: Model Selection Criteria: Prater Data

Figure 1: Residuals versus fitted values for simple linear regression on Hooker data.

Figure 2: y versus quadratic model fitted values on Hooker data.

Figure 3: y versus exponentiated fitted values of simple linear regression on $\log(y)$ for Hooker data.

Figure 4: Residuals versus $(x_i - \bar{x})^2$ for for simple linear regression on Hooker data.