A USEFUL INTERPRETATION OF $R^2$ IN BINARY CHOICE MODELS
(OR, HAVE WE DISMISSED THE GOOD OLD $R^2$ PREMATURELY)

Reuben Gronau
Industrial Relations Section
Princeton University
Hebrew University, Jerusalem

January 1998
A USEFUL INTERPRETATION OF $R^2$ IN BINARY CHOICE MODELS
(OR, HAVE WE DISMISSED THE GOOD OLD $R^2$ PREMATURELY)

Abstract

The discreditation of the Linear Probability Model (LPM) has led to the dismissal of the standard $R^2$ as a measure of goodness-of-fit in binary choice models. It is argued that as a descriptive tool the standard $R^2$ is still superior to the measures currently in use.

In the LPM model $R^2$ has a simple interpretation: it equals the difference between the average predicted probability in the two groups. It also measures the fraction of the explained part of the variance (SSR) due to the difference between the conditional means (SSB). Given $R^2$ and the sample proportion $P$ one can calculate the conditional means, $\bar{P}_0$ and $\bar{P}_1$.

This interpretation still holds for non-linear cases when $R^2$ is computed as the regression coefficient of the predicted value on the dependent binary variable. However, even if other definitions of $R^2$ are used in this case (e.g., the share of the variance explained by the regression, or the correlation coefficient between true and predicted values), the measure is very close to $\bar{P}_1 - \bar{P}_0$. 
A USEFUL INTERPRETATION OF R² IN BINARY CHOICE MODELS
(OR, HAVE WE DISMISSED THE GOOD OLD R² PREMATURELY)'

The linear probability model (LPM) was discredited in the econometric literature long ago. Heteroskedasticity and, more important, the fact that its predicted values are not confined to the range [0,1], resulted in the abandonment of the straight line in favor of its more glamorous S-shaped relatives. Nowadays most econometric textbooks mention LPM only in passing, as an introduction to logit, probit and more sophisticated multi-choice models.

An innocent casualty of the loss of faith in LPM was the standard measure of goodness-of-fit, R², which was superseded by a series of "pseudo-R²" based on the likelihood ratio. I shall argue that by doing so we lost a useful descriptive tool describing the predictive quality of our models.

My starting point is the LPM. Let the population consist of two groups A₀ and A₁. To determine the choice of A₁, a linear regression is fitted, \( Y = Xb + e \), where \( Y_i \) is a binary variable,

\[
Y_i = \begin{cases} 
1 & \text{if } i \in A_1 \\
0 & \text{if } i \in A_0 
\end{cases}
\]

\( X \) is a matrix of explanatory variables and \( e_i \) is the residual. In the linear model, \( \hat{Y}_i = P_i = Xb \) is the predicted probability that observation \( i \) belongs to \( A_1 \).

The LPM employs OLS and, hence (assuming \( X \) includes a constant term) \( \bar{e} = 0 \), \( \bar{P} = P \) and \( \sum xe = \sum pe = 0 \), where \( \bar{P} \) denotes the average predicted value, \( P \) is the sample mean of \( Y \) (\( P = n_i/n \), where \( n_i \) is the number of observations in the sample belonging to group \( A_i \)), and
lower-case letters denote deviations from the mean. The coefficient of determination $R^2$
measures, as in the continuous variable case, the ratio of the variances of the predicted
probability and of $Y$ (Goldberger, 1973):

\[ (1) \quad R^2 = \frac{V(P)/V(Y)}{= \frac{(\sum p y)^2}{\sum p^2 \sum y^2} = \frac{\sum p y}{\sum y^2}} = \frac{[\sum Y_i P_i - (\sum Y_i \sum P_i/n)/[\sum Y_i^2 - (\sum Y_i^2)/n]]}{-} \]

Since $\sum Y_i = \sum P_i = \sum Y_i^2 = n_i$, and $\sum Y_i P_i = n_i \bar{P}_i$, where $\bar{P}_i$ is the mean predicted
probability for the observations belonging to $A_i$,

\[ (2) \quad R^2 = \frac{n_i \bar{P}_i - (n_i^2/n)}{[n_i - (n_i^2/n)]} = \frac{(\bar{P}_i - P)/(1-P)}{(1-P) \bar{P}_i} \]

Equation (2) has a simple interpretation: $P$ is the naive predictor (in the absence of knowledge
of $X$), thus the denominator describes the error of using this predictor for group $A_i$. The
numerator $\bar{P}_i - P$ indicates the improvement offered by the regression. Hence $R^2$ measures the
relative improvement of prediction for group $A_i$.

Similarly, since $P = \bar{P} = P \bar{P}_i + (1-P) \bar{P}_o$,

\[ (3) \quad \frac{(1-\bar{P})}{(1-P)} = \bar{P}_o / P, \]
and (2) can be rewritten as:

\[ (4) \quad R^2 = 1 - [(1-\bar{P})/(1-P)] = (P - \bar{P}_o)/P, \]
which has the same interpretation when applied to $A_o$. 2
Given \( P \) and \( R^2 \) one can compute, using (2) and (4), the conditional sample means

\[
\bar{P}_1 = P + R^2 (1 - P) = P(1 - R^2) + R^2
\]

and

\[
\bar{P}_0 = P - PR^2 = P(1 - R^2).
\]

Comparing (5) with (6) yields the surprising result

\[
R^2 = \bar{P}_1 - \bar{P}_0.
\]

\( R^2 \) is the difference between the conditional means. In the extreme cases:

a. The model is a perfect predictor, \( \bar{P}_1 = 1 \) , \( \bar{P}_0 = 0 \), and \( R^2 = 1 \);

b. LPM does not improve the prediction relative to the naive model, \( \bar{P}_0 = \bar{P}_1 = P \),

and hence \( R^2 = 0 \).

Alternatively,

\[
1 - R^2 = \bar{P}_0 + (1 - \bar{P}_1).
\]

\( \bar{P}_0 \) is the mean probability of an error for group \( A_0 \) (i.e., the probability of predicting \( Y_i = 1 \) when \( Y_i = 0 \)), \( 1 - \bar{P}_1 \) is the same probability for group \( A_1 \) (predicting \( Y_i = 0 \) when \( Y_i = 1 \)), hence \( 1 - R^2 \) measures the sum of the probabilities of committing these errors.

Additional light ensues when the analysis is cast in the ANOVA framework. Total variation \( \sum y^2 \) can be broken into three parts: the part explained by between-group variation in prediction, the part explained by within-group variation and the unexplained residual.

Recalling that \( \sum P_i e_i = \sum P_i e_i = \sum P_i (Y_i - \bar{P}) = 0 \)
\begin{align}
(9) \quad \sum P_i(Y_i - \bar{P}_i) = \sum P_i - \sum P_i^2 = n\bar{P}_1 - \sum P_i^2 = 0
\end{align}

Hence, using (9) and (3),

\begin{align}
(10) \quad SSE = \sum e_i^2 = \sum (Y_i - P_i)^2 = \sum Y_i - 2\sum Y_iP_i + \sum P_i^2 = n(1 - 2\bar{P}_1) + \sum P_i^2 = n(1 - \bar{P}_1) = n_0\bar{P}_0
\end{align}

The sum of the squared residuals equals the number of observations in \( A_1 \) which are predicted to belong to \( A_0 \) (and vice versa).

The between-group variation equals

\begin{align}
(11) \quad SSB = \sum \frac{1}{n_i}(\bar{P}_i - \bar{P})^2 = nP(1 - P)(\bar{P}_1 - \bar{P}_0)^2 = n \sigma(Y) R^4.
\end{align}

Hence the share of the explained variation due to between-group variation equals \( R^2 \), within-group variation contributing the rest. For low \( R^2 \) most of the explanatory power of the regression rests in the within-group variation. It is only when \( R^2 > 0.5 \) that the discriminating power of the regression (i.e., the between-group variation) becomes the dominant component, and the within-group variation diminishes in both relative and absolute terms. Collecting the results and rearranging yields the ANOVA table (Table 1).

Can these results be generalized to other estimation methods of binary choice? The underlying assumptions that generate these results are those embodied in the OLS normal equations (\( \bar{e} = 0 \) and \( \sum xe = 0 \)), and the assumption of linearity (\( P = Xb \) and hence \( \sum pe = 0 \)).

Even if the non-linear models pass the first test (e.g., the logit) they fail (by definition) the second. In this case one has to amend the variance decomposition formula

\begin{align}
(12) \quad SST = SSE + SSR + n(\bar{P} - P)^2 = 2S_n,
\end{align}

4
Table 1: Analysis of Variance of the Binary Variable

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sum of squares</th>
<th>Share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explained by LPM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>of which:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between-groups</td>
<td>( SSR = \sum_{j=0}^{i} n_j (\bar{P}_j - \bar{P})^2 = nP(1-P)(\bar{P}_1 - \bar{P}_0)^2 )</td>
<td>( R^2 = (\bar{P}_1 - \bar{P}_0)^2 )</td>
</tr>
<tr>
<td>Within-groups</td>
<td>( SSW = \sum_{j=0}^{i} \sum_{r=1}^{n} (P_r - \bar{P}_j)^2 = nP(1-P)(\bar{P}_1 - \bar{P}_0)(1 - (\bar{P}_1 - \bar{P}_0)) )</td>
<td>( R^2(1-R^2) )</td>
</tr>
<tr>
<td>Unexplained</td>
<td>( SSE = \sum e_i^2 = n(1-P)(\bar{P}_1 - \bar{P}_0)^2 = nP(1-P)[(1-\bar{P}_1) + \bar{P}_0] )</td>
<td>( 1 - R^2 = (1-\bar{P}_1) + \bar{P}_0 )</td>
</tr>
<tr>
<td>Total</td>
<td>( SST = \sum y_i^2 = nP(1-P) )</td>
<td>( 1 )</td>
</tr>
</tbody>
</table>
where SSR = SSB + SSW and $S_{\sigma}$ is the covariance between the error and the predictor

$$ (13) \quad S_{\sigma} = \sum \sum (Y_0 - P_0) P_0 - n \bar{P} (P - \bar{P}). $$

In this case the correlation coefficient between the dependent variable and the predicted value

$$ R^2 = (\sum \gamma P / n)^2 / V(P) \cdot V(Y) $$

does not equal the ratio of the variances of the predicted and the true values $R^2 = V(P)/V(Y) = SSR/SST$. However,

$$ (14) \quad \frac{\sum \gamma P}{n} = \frac{n \sum Y_i P_i - (\sum Y_0 \sum P_i) / n = n P(1 - P) (\bar{P}_1 - \bar{P}_0) }{(\bar{P}_1 - \bar{P}_0)n V(Y)} .$$

Hence, by (14) and (11)

$$ (15) \quad R^2 = SSB/SSR = (\bar{P}_1 - \bar{P}_0)^2 / R^2, $$

and the geometric mean of $R^2$ and $R^2$ equals

$$ (16) \quad R^2 = R_1 R_2 = SSB/SST = \frac{\sum \gamma P}{\sum \gamma Y} = \frac{\bar{P}_1 - \bar{P}_0}{R^2} . $$

The regression coefficient of $P$ on $Y$ equals the difference between the predicted means of the two subsamples, and can still be used as a measure of the goodness of fit in the non-linear case.

The difference between $R^2$ and the other measures $R^2$ and $R^2$ depends on the prediction bias ($\bar{P} - P$) and on the covariance $S_{\sigma}$. The extent of the "damage" caused by these two factors cannot be evaluated without a detailed Monte-Carlo experiment, but judging from a limited experiment it seems that the share of these two factors in total variation is small.

Table 2 contains the decomposition of the variance for two cases: the dependent variable in the first is car ownership, and in the second - the labor force participation (LFP) of women. In both cases the data used are the Israeli Time Use Survey. Since there is no interest in the results proper, they are suppressed, and the table reports only the share of total variation of the
different components using LPM, logit and probit. The explanatory power in the two experiments is substantially different ($R^2 = 0.28$ for cars and 0.18 for LFP, using LPM), and in the car-ownership case the non-linear methods seem to fit the data better. In both cases the "noise" introduced by the covariance and the prediction bias is minimal, and the three measures of $R^2$ are very close also in the logit and probit cases.²

A whole array of statistics is currently used to measure the goodness of fit of the non-linear regressions. How do these measures relate to the standard $R^2$? All these measures are based on the comparison of the value of the maximized likelihood function $L_{cr}$ with the value of the function under the assumption that the explanatory variables do not affect choice $L_n$ (i.e., under the restriction $b = 0$). The value of the maximized likelihood function equals

$$L_{cr} = \prod_{i=1}^{n} (1-F(-x\hat{b}))^n [1-F(-x\hat{b})]^{1-n} = \prod_{i=1}^{n} P_i^b (1-P_i)^{1-b}$$

where $F(\ )$ is the distribution function and $\hat{b}$ are the non-linear estimates $P_i = 1 - F(-x\hat{b})$.

Rearranging yields

$$L_{cr} = P_i^b (1-P_i)^{1-b}$$

(17)  $$L_{cr} = \tilde{P}_i^b (1-\tilde{P}_i)^{1-b} = \exp \{n[\lnP_i^b(1-P_i)^{1-b}]\}$$

where $\tilde{P}_i$ and $(1-\tilde{P}_i)$ are the geometric means of $P_i$ of the two groups.³ Similarly, $L_n$ is estimated under the assumption that the predictor is the naive predictor $P_i = P$, and hence

(18)  $$L_n = P^b (1-P)^{1-b} = \exp \{n[\lnP(1-P)^{1-b}]\}$$

Note that both (17) and (18) are variants of a measure of entropy, but that $L_{cr}$ is based on the geometric rather than on the arithmetic mean.
Table 2: The Decomposition of Variance

<table>
<thead>
<tr>
<th>Estimation method</th>
<th>Car ownership</th>
<th>LFP of women</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LPM</td>
<td>Logit</td>
<td>Probit</td>
</tr>
<tr>
<td>n</td>
<td>2426</td>
<td>2426</td>
<td>2426</td>
</tr>
<tr>
<td>P</td>
<td>0.59100</td>
<td>0.59100</td>
<td>0.59100</td>
</tr>
<tr>
<td>\bar{P}</td>
<td>0.59100</td>
<td>0.59100</td>
<td>0.58965</td>
</tr>
<tr>
<td>\bar{e}</td>
<td>-</td>
<td>-</td>
<td>0.00145</td>
</tr>
<tr>
<td>\bar{P}_1</td>
<td>0.70735</td>
<td>0.73443</td>
<td>0.72942</td>
</tr>
<tr>
<td>\bar{P}_0</td>
<td>0.42303</td>
<td>0.38390</td>
<td>0.38760</td>
</tr>
<tr>
<td>\bar{P}_1 - \bar{P}_0</td>
<td>0.28432</td>
<td>0.35033</td>
<td>0.34182</td>
</tr>
</tbody>
</table>

SHARES

<table>
<thead>
<tr>
<th></th>
<th>LFP of women</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR</td>
<td>0.28432</td>
<td>0.34470</td>
</tr>
<tr>
<td>SSW</td>
<td>0.20348</td>
<td>0.22182</td>
</tr>
<tr>
<td>SSB</td>
<td>0.08084</td>
<td>0.12288</td>
</tr>
<tr>
<td>SSE</td>
<td>0.71568</td>
<td>0.64363</td>
</tr>
<tr>
<td>n(\bar{P} - \bar{P})^2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2S_e</td>
<td>-</td>
<td>0.01166</td>
</tr>
<tr>
<td>SST</td>
<td>1.00000</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SSB/SSR</td>
<td>0.28432</td>
</tr>
<tr>
<td>R^2_1</td>
<td>*</td>
</tr>
</tbody>
</table>

* Cannot be computed (see text).
Several combinations of the ratio of $L_{UR}$ and $L_R$ have been suggested as measures of goodness-of-fit in the non-linear case. The most common (Maddala, 1983) are

$$\tilde{R}_i^2 = 1 - \left( \ln L_{UR} / \ln L_R \right)$$

(19) $$\tilde{R}_i^2 = 1 - \left( L_{UR} / L_R \right)^{2n}$$

$$\tilde{R}_i^2 = \tilde{R}_i^2 / \left( 1 - L_{UR}^{2n} \right).$$

Inserting the value of $L_{UR}$ and $L_R$ into these terms yields

(20a) $$\tilde{R}_i^2 = 1 - \left[ \ln (\tilde{P}_i / P) + (1 - P) \ln (1 - P) \right] / \left[ \ln P + (1 - P) \ln (1 - P) \right]$$

$$= - \left[ \ln (\tilde{P}_i / P) + (1 - P) \ln \left( \left( \frac{1 - P_i}{1 - P} \right)^{1 - P} \right) \right] / \left[ \ln P + (1 - P) \ln (1 - P) \right]$$

(20b) $$\tilde{R}_i^2 = 1 - \left( \frac{\tilde{P}_i}{P} \right) \left( \frac{(1 - P_i)}{(1 - P)} \right)^{1 - P}$$

(20c) $$\tilde{R}_i^2 = \tilde{R}_i^2 / \left( 1 - P \left( 1 - P_i \right)^{1 - P} \right)^2.$$

By comparison, combining (2) and (4) $R^2$ is approximately

(21) $$R^2 = \left( \frac{P_i}{P} \right) \left[ \left( \frac{1 - P_i}{1 - P} \right)^{1 - P} \right] / \left( 1 - P \right)$$

All four measures reflect the improved prediction due to the regression comparing the groups' mean predicted values with the naive predictors. The current measures (20) and $R^2$ differ in two respects:

a. The current measures use the geometric group mean, while $R^2$ uses the arithmetic mean,

and

b. The current measures weigh the improvement in prediction in the two groups differently (the weight depending on $P$).
The bottom of Table 2 contains the estimates of $\hat{R}_1^2$ for the logit and the cases. (Since $\hat{R}_1^2$ is based on a geometric mean, it cannot be applied to LPM when LPM generates negative predictions.) The estimates are substantially lower than $R^2$. This is most probably the result of using a geometric rather than an arithmetic mean.

There is no prima facie advantage in using the geometric mean or the measure of entropy, rather than the arithmetic mean and the variance, especially since $P$ is the sample arithmetic mean. In his survey of the different measures of goodness of fit in qualitative response (QR) models Amemiya (1981, p. 1504) argues that the use of $R^2$ in this context "cannot be defended as strongly as in the standard regression model because a QR model is essentially a heteroscedastic regression model." This note shows, however, that as a descriptive tool $R^2$ is still superior to all other measures. Given the rich interpretation that $R^2$ assumes in the binary variable case, it seems that its dismissal was premature, and that it should be reinstated in the standard econometric tool-kit used to analyze qualitative choice.
References:


This note was instigated by a home assignment in Ernst Berndt, *The Practice of Econometrics*. He is, however, completely innocent of the outcome. I benefitted from the comments of Arthur Goldberger and participants of the workshops at the Industrial Relations Section at Princeton and at the Hebrew University. Arthur Goldberger drew my attention to the fact that some of my results are implicit in Ladd (1966). Ladd, however, focused on the similarity of the regression coefficients in linear probability functions and discriminant analysis, and barely touches on the measures of goodness of fit.

1. It is comforting to know that since $R^2 \geq 0$ LPM will never produce the embarrassing result $\hat{P}_0 > \hat{P}_1$. Equations (4) and (5) also guard us from the equally embarrassing results $P_1 > 1$ and $P_0 < 0$.

2. In a dozen experiments that I conducted the prediction bias is infinitely small and the covariance term $(2S_{20})$ never exceeded 2 percent of the total variation. As a result the error of using $\hat{P}_1 - \hat{P}_0$ as a measure of $R^2$ is in the range of $\pm 0.008$.

3. Note that $(1 - P_0) \neq 1 - \hat{P}_0$.

4. There is no agreement concerning the names of the different measures. Maddala calls $\tilde{R}_i$ the McFadden’s $R^2$ and $\tilde{R}_i$ the "pseudo $R^2$". The STATA computer program computes $\tilde{R}_i$ as the "pseudo $R^2$". The three measures are, of course, related, since

$$\tilde{R}_i = 1 - L_{R_i^{20}} = 1 - \{P^T (1-P)^T \}^{20}.$$

5. Equation (20) assumes that the prediction bias is negligible $P = \tilde{P}$, an assumption borne out by our observations.