4 Hypothesis Testing and Goodness of Fit

This chapter begins by reviewing tests of hypothesis that can be used with any model estimated by maximum likelihood. Next, methods for detecting outliers and influential observations for the binary logit and probit models are examined; comparable methods for ordinal and nominal outcomes are not available. The chapter ends with a review of scalar measures for assessing the overall goodness of fit of a model. While some of these measures apply only to the binary response model, most can be adapted to the models in later chapters.

4.1. Hypothesis Testing

ML estimators are distributed asymptotically normally. This means that as the sample size increases, the sampling distribution of an ML estimator becomes approximately normal. For an individual parameter,

$$\widehat{\beta}_k \overset{a}{\sim} \mathcal{N}(\beta_k, \text{Var}(\widehat{\beta}_k))$$

where "$\overset{a}{\sim}$" reads "is distributed asymptotically as." For a vector of parameters,

$$\widehat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, \text{Var}(\widehat{\boldsymbol{\beta}}))$$

where $\text{Var}(\widehat{\boldsymbol{\beta}})$ is the covariance matrix for $\widehat{\boldsymbol{\beta}}$. For example, with three coefficients:

$$\text{Var}\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\\ \widehat{\beta}_2\end{pmatrix} = \begin{pmatrix}\sigma^2_{\widehat{\beta}_0} & \sigma_{\widehat{\beta}_0,\widehat{\beta}_1} & \sigma_{\widehat{\beta}_0,\widehat{\beta}_2}\\ \sigma_{\widehat{\beta}_1,\widehat{\beta}_0} & \sigma^2_{\widehat{\beta}_1} & \sigma_{\widehat{\beta}_1,\widehat{\beta}_2}\\ \sigma_{\widehat{\beta}_2,\widehat{\beta}_0} & \sigma_{\widehat{\beta}_2,\widehat{\beta}_1} & \sigma^2_{\widehat{\beta}_2}\end{pmatrix}$$

The off-diagonal elements are the covariances between the estimates of two parameters.

Consider the simple hypothesis H_0: $\beta_k = \beta^*$, where β^* is the hypothesized value, often equal to 0. Since $\sigma_{\widehat{\beta}_k}$ is unknown, it must be estimated, which results in the test:

$$z = \frac{\widehat{\beta}_k - \beta^*}{\widehat{\sigma}_{\widehat{\beta}_k}} \qquad [4.1]$$

Under the assumptions justifying ML, if H_0 is true, then z is distributed approximately normally with a mean of 0 and a variance of 1 for large samples. The sampling distribution for z, drawn in Figure 4.1, shows the probability of various values of z when H_0 is true. For example, the shaded region for $z > 1.96$ indicates that values of z greater than 1.96 will occur due to sampling variation 2.5% of the time. Similarly, the shaded region on the left indicates how frequently values less than -1.96 will occur. For a two-tailed test, H_0 is rejected at the .05 level when z falls in the shaded region of either tail. If past research or theory suggests the sign of the coefficient, a one-tailed test is used and the null hypothesis would only be rejected when z is in the expected tail.

The test statistic in Equation 4.1 is sometimes considered to have an asymptotic t-distribution, and the test is referred to as a t-test or a quasi-t-test. When N is large, which is required for the asymptotic justification of the test, it makes little difference whether a t-distribution or a normal

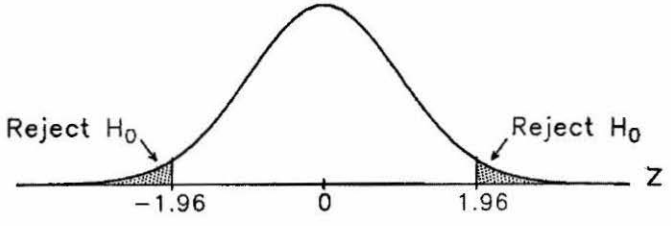

Figure 4.1. Sampling Distribution for a z-Statistic

distribution is used. Accordingly, some programs label this statistic a z-test, while other programs label it a t-test.

Example of the z-Test: Labor Force Participation

To test the hypothesis that having young children affects a woman's probability of working, we can use the z-statistic in Table 3.3 for the logit model. Since prior research suggests that the effect is negative, a one-tailed test is used. We conclude that:

- Having young children has a significant effect on the probability of working ($z = -7.43$, $p < .01$ for a one-tailed test).

4.1.1. Wald, Likelihood Ratio, and Lagrange Multiplier Tests

It is often useful to test complex hypotheses. For example, you might want to test that several coefficients are simultaneously equal to 0, or that two coefficients are equal. Such hypotheses can be tested with Wald, likelihood ratio (LR), or Lagrange multiplier (LM) tests. These tests can be thought of as a comparison between the estimates obtained after the constraints implied by the hypothesis have been imposed to the estimates obtained without the constraints. This is illustrated in panel A of Figure 4.2, which is based on a figure from Buse (1982).

The log likelihood function for estimating β is drawn as a solid curve. The *unconstrained* estimator $\widehat{\beta}_U$ maximizes the log likelihood function, with the log likelihood equal to $\ln L(\widehat{\beta}_U)$. The hypothesis H_0: $\beta = \beta^*$ imposes the constraint $\beta = \beta^*$, so that the *constrained* estimate $\widehat{\beta}_C$ equals β^*. Unless $\widehat{\beta}_U$ is *exactly* equal to β^*, $\ln L(\widehat{\beta}_C)$ is smaller than $\ln L(\widehat{\beta}_U)$, as shown in the figure. The LR test assesses the constraint by comparing the log likelihood of the unconstrained model, $\ln L(\widehat{\beta}_U)$, to the log likelihood of the constrained model, $\ln L(\widehat{\beta}_C)$. If the constraint significantly reduces the likelihood, then the null hypothesis is rejected.

The Wald test estimates the model without constraints, and assesses the constraint by considering two things. First, it measures the distance between the unconstrained and the constrained estimates. In our example, this quantity is $\widehat{\beta}_U - \widehat{\beta}_C = \widehat{\beta}_U - \beta^*$. The larger the distance, the less likely it is that the constraint is true. Second, this distance $\widehat{\beta}_U - \widehat{\beta}_C$ is weighted by the curvature of the log likelihood function, which is indicated by the second derivative $\partial^2 \ln L/\partial\beta^2$. The larger the second derivative, the faster the curve is changing. (*What does it mean if the second derivative is* 0?) The importance of the shape of the function is illustrated in panel B. The log likelihood drawn with a dashed line is nearly

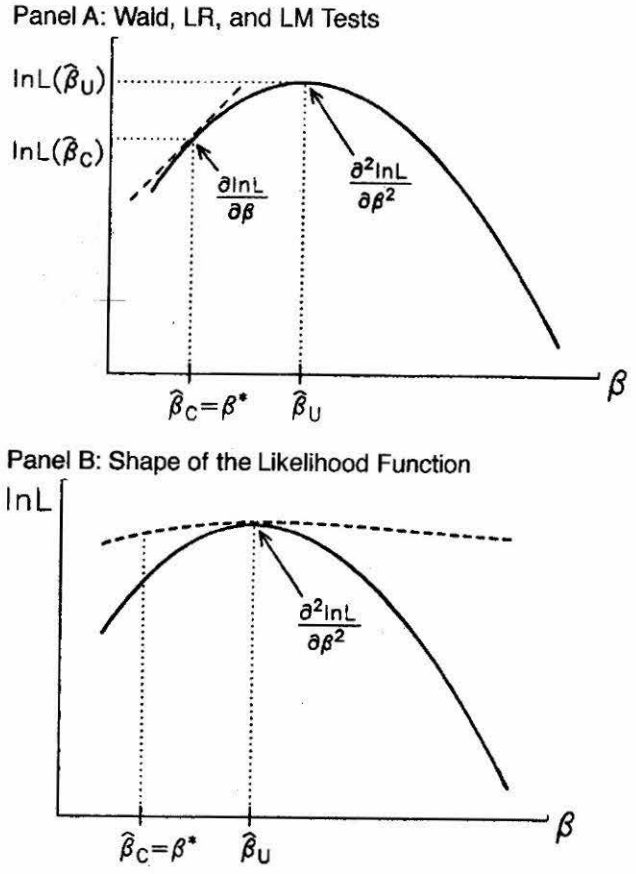

Figure 4.2. Wald, Likelihood Ratio, and Lagrange Multiplier Tests

flat, so the second derivative evaluated at $\widehat{\beta}_U$ is relatively small. When the second derivative is small, the distance between $\widehat{\beta}_U$ and $\widehat{\beta}_C$ is minor relative to the sampling variation. The second function, drawn with a solid line, has a larger second derivative, indicating a more rapidly changing likelihood function. With a larger second derivative, the same distance between $\widehat{\beta}_U$ and $\widehat{\beta}_C$ might be significant. (*How would increasing the sample size affect the curvature of the log likelihood function?*)

The Lagrange multiplier (LM) test, also known as the score test, only estimates the constrained model, and assesses the slope of the log likelihood function at the constraint. If the hypothesis is true, the slope (known as the score) at the constraint should be close to 0. In panel A of Figure 4.2, the slope is represented by the tangent to the curve drawn with a dashed line, which is labeled $\partial \ln L/\partial\beta$. As with the Wald test, the curvature of the log likelihood function at the constraint is used to assess the significance of a nonzero slope.

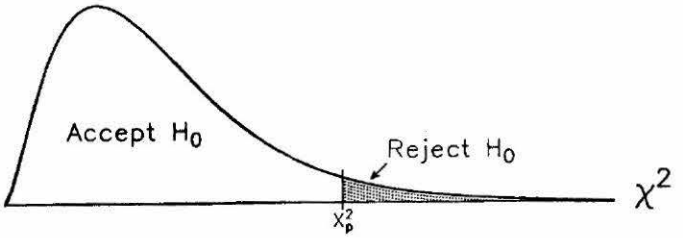

Figure 4.3. Sampling Distribution of a Chi-Square Statistic with 5 Degrees of Freedom

When H_0 is true, the Wald, LR, and LM tests are *asymptotically* equivalent. As N increases, the sampling distributions of the three tests converge to the same chi-square distribution with degrees of freedom equal to the number of constraints being tested. Figure 4.3 shows the sampling distribution for a chi-square statistic with 5 degrees of freedom. The area to the right of X^2_p is equal to p, and indicates the probability of observing a value of the test statistic greater than X^2_p if H_0 is true. The null hypothesis is rejected at the p level of significance if the test statistic is larger than X^2_p.

It is important to remember that the Wald, LR, and LM tests only have asymptotic justifications. The degree to which these tests approximate a chi-square distribution in small samples is largely unknown. See Section 3.5.1 (p. 53) for guidelines on the sample size needed for using these tests.

With these ideas in mind, we are ready for formal definitions of the Wald and LR tests. The LM test is discussed further in Chapter 7.

4.1.2. The Wald Test

While in its most general form the Wald test can be used to test nonlinear constraints, here we consider only linear constraints of the form:

$$\mathbf{Q}\boldsymbol{\beta} = \mathbf{r} \qquad [4.2]$$

where $\boldsymbol{\beta}$ is the vector of parameters being tested, $\mathbf{Q}$ is a matrix of constants, and $\mathbf{r}$ is a vector of constants. While we are usually interested only in the intercepts and slopes of a model, $\boldsymbol{\beta}$ could contain other parameters such as σ in the LRM. By specifying $\mathbf{Q}$ and $\mathbf{r}$, a variety of linear constraints can be imposed. For example, consider the probit model $\Pr(y = 1 \mid \mathbf{x}) = \Phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$. To test that $\beta_1 = 0$, Equation 4.2 becomes

$$(0\,1\,0)\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix} = (0)$$

Or, to test the constraint that $\beta_1 = \beta_2 = 0$,

$$\begin{pmatrix}0\,1\,0\\ 0\,0\,1\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}$$

The hypothesis H_0: $\mathbf{Q}\boldsymbol{\beta} = \mathbf{r}$ can be tested with the Wald statistic:

$$W = [\mathbf{Q}\widehat{\boldsymbol{\beta}} - \mathbf{r}]'[\mathbf{Q}\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\mathbf{Q}']^{-1}[\mathbf{Q}\widehat{\boldsymbol{\beta}} - \mathbf{r}] \qquad [4.3]$$

W is distributed as chi-square with degrees of freedom equal to the number of constraints (i.e., the number of rows of $\mathbf{Q}$). The Wald statistic consists of two components. First, $\mathbf{Q}\widehat{\boldsymbol{\beta}} - \mathbf{r}$ at each end of the formula measures the distance between the estimated and hypothesized values. Second, $[\mathbf{Q}\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\mathbf{Q}']^{-1}$ reflects the variability in the estimator, or, alternatively, the curvature of the likelihood function. To see this more clearly, consider a simple example.

For the model $\Pr(y = 1 \mid \mathbf{x}) = \Phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ with H_0: $\beta_1 = \beta^*$, $\mathbf{Q}\widehat{\boldsymbol{\beta}} - \mathbf{r}$ can be written as

$$(0\,1\,0)\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\beta}_1\\ \widehat{\beta}_2\end{pmatrix} - (\beta^*) = \widehat{\beta}_1 - \beta^*$$

$\mathbf{Q}\widehat{\boldsymbol{\beta}} - \mathbf{r}$ is repeated at the end of the formula, which squares the distance between the hypothesized value and the estimate. Therefore, negative and positive distances have the same effect on the test statistic. The middle portion of the Wald statistic is

$$[\mathbf{Q}\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\mathbf{Q}']^{-1} = \left[(0\,1\,0)\,\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\begin{pmatrix}0\\ 1\\ 0\end{pmatrix}\right]^{-1} = \frac{1}{\widehat{\sigma}^2_{\widehat{\beta}_1}}$$

which is simply the inverse of the variance. The larger the variance, the smaller the weight given to the distance between the hypothesized and estimated value. Equivalently, the faster the likelihood function is changing in the region around $\widehat{\beta}_1$, the more significant the difference $\widehat{\beta}_1 - \beta^*$. (*Why should we give less weight when the variance is larger?*) Combining these results,

$$W = \frac{(\widehat{\beta}_1 - \beta^*)^2}{\widehat{\sigma}^2_{\widehat{\beta}_1}} = \left(\frac{\widehat{\beta}_1 - \beta^*}{\widehat{\sigma}_{\widehat{\beta}_1}}\right)^2$$

which is distributed as chi-square with 1 degree of freedom if H_0 is true. Notice that W is the square of the z-statistic in Equation 4.1, which corresponds to a chi-square variable with 1 degree of freedom being equal to the square of a normal variable. Some programs, such as SAS, present a single degree of freedom chi-square statistic for individual coefficients, rather than the z-statistic.

The same ideas apply to more complex hypotheses. Consider H_0: $\beta_1 = \beta_2 = 0$, which can be written as

$$H_0\text{: }\begin{pmatrix}0\,1\,0\\ 0\,0\,1\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}$$

$\mathbf{Q}\widehat{\boldsymbol{\beta}} - \mathbf{r}$ is simply $(\widehat{\beta}_1\ \widehat{\beta}_2)'$. The middle portion of the Wald formula is

$$[\mathbf{Q}\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\mathbf{Q}']^{-1} = \left[\begin{pmatrix}0\,1\,0\\ 0\,0\,1\end{pmatrix}\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\begin{pmatrix}0\,0\\ 1\,0\\ 0\,1\end{pmatrix}\right]^{-1}$$

To keep the example simple, assume that the estimates are uncorrelated. (In practice, the estimates will be correlated.) Then

$$[\mathbf{Q}\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\mathbf{Q}']^{-1} = \begin{pmatrix}\widehat{\sigma}^2_{\widehat{\beta}_1} & 0\\ 0 & \widehat{\sigma}^2_{\widehat{\beta}_2}\end{pmatrix}^{-1} = \begin{pmatrix}1/\widehat{\sigma}^2_{\widehat{\beta}_1} & 0\\ 0 & 1/\widehat{\sigma}^2_{\widehat{\beta}_2}\end{pmatrix} \qquad [4.4]$$

The larger the variance, the less weight is given to the distance between the hypothesized and estimated parameter. Carrying out the algebra, we obtain

$$W = \sum_{k=1}^{2}\frac{\widehat{\beta}^2_k}{\widehat{\sigma}^2_{\widehat{\beta}_k}} = \sum_{k=1}^{2}\left(\frac{\widehat{\beta}_k}{\widehat{\sigma}^2_{\widehat{\beta}_k}}\right)^2 = \sum_{k=1}^{2} z^2_{\widehat{\beta}_k}$$

With uncorrelated parameters, the Wald statistic is the sum of squared z's. Recall that a chi-square distribution with J degrees of freedom is defined as the sum of J independent, squared normal random variables. When the estimates are correlated, which is normally the case, the resulting formula is more complicated, but the general ideas are the same.

Examples of the Wald Test: Labor Force Participation

To illustrate the Wald test, consider the logit model:

$$\Pr(LFP = 1) = \Lambda(\beta_0 + \beta_1 K5 + \beta_2 K618 + \beta_3 AGE + \beta_4 WC + \beta_5 HC + \beta_6 LWG + \beta_7 INC) \quad [4.5]$$

Wald Test of a Single Coefficient. To test H_0: $\beta_1 = 0$, let

$$\mathbf{Q} = (0\ 1\ 0\ 0\ 0\ 0\ 0\ 0) \qquad \text{and} \qquad \mathbf{r} = (0)$$

Then $W = 55.14$, which is the square of the z-statistic for $K5$ in Table 3.3. We describe the result as:

- The effect of having young children on the probability of entering the labor force is significant at the .01 level ($X^2 = 55.14$, $df = 1$, $p < .01$).

The symbol X^2 is often used rather than W since the Wald statistic has a chi-square distribution.

Wald Test That Two Coefficients Are 0. The hypothesis that the effects of the husband's and wife's education are simultaneously 0 can be written as: H_0: $\beta_4 = \beta_5 = 0$. To test this hypothesis, let

$$\mathbf{Q} = \begin{pmatrix} 0\,0\,0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,0\,0\,1\,0\,0 \end{pmatrix} \qquad \text{and} \qquad \mathbf{r} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Then $W = 17.66$ with 2 degrees of freedom. We conclude:

- The hypothesis that the effects of the husband's and wife's education are simultaneously equal to 0 can be rejected at the .01 level ($X^2 = 17.66$, $df = 2$, $p < .01$).

(*Specify* **Q** *and* **r** *to test the hypothesis that all of the coefficients except the intercept are simultaneously* 0.)

Wald Test That Two Coefficients Are Equal. To test that the effect of the husband's education equals the effect of the wife's education, define

$$\mathbf{Q} = (0\ 0\ 0\ 0\ 1\ {-1}\ 0\ 0) \qquad \text{and} \qquad \mathbf{r} = (0)$$

Substituting these matrices into Equation 4.3 and simplifying results in the usual formula:

$$W = \frac{(\hat{\beta}_4 - \hat{\beta}_5)^2}{\widehat{\text{Var}}(\hat{\beta}_4) + \widehat{\text{Var}}(\hat{\beta}_5) - 2\widehat{\text{Cov}}(\hat{\beta}_4, \hat{\beta}_5)}$$

Then $W = 3.54$ with 1 degree of freedom. There is 1 degree of freedom since there is a single restriction, even though that restriction involves two parameters. We conclude:

- The hypothesis that the effects of the husband's and wife's education are equal is marginally significant at the .05 level ($X^2 = 3.54$, $df = 1$, $p = .06$).

4.1.3. The Likelihood Ratio Test

The LR test can also be used to test constraints on a model. While in its most general form these constraints can be complex and nonlinear, I only consider constraints that involve eliminating one or more regressors from the model. For example, consider the logit models:

$$M_1: \quad \Pr(y = 1 \mid \mathbf{x}) = \Lambda(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$
$$M_2: \quad \Pr(y = 1 \mid \mathbf{x}) = \Lambda(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$$
$$M_3: \quad \Pr(y = 1 \mid \mathbf{x}) = \Lambda(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4)$$
$$M_4: \quad \Pr(y = 1 \mid \mathbf{x}) = \Lambda(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)$$

Model M_1 is formed from M_2 by imposing the constraint $\beta_3 = 0$, and M_1 is formed from M_3 by imposing the constraint $\beta_4 = 0$. When one model can be obtained from another model by imposing constraints, the *constrained* model is said to be *nested* in the *unconstrained* model. Thus, M_1 is nested in M_2 and in M_3 . However, M_2 is not nested in M_3, nor is M_3 nested in M_2. (*Which models are nested in* M_4*?*)

The LR test is defined as follows. The constrained model M_C with parameters $\boldsymbol{\beta}_C$ is nested in the unconstrained model M_U with parameters $\boldsymbol{\beta}_U$. The null hypothesis is that the constraints imposed to create M_C are true. Let $L(M_U)$ be the value of the likelihood function evaluated at the ML estimates for the unconstrained model, and let $L(M_C)$ be the value

at the constrained estimates. The *likelihood ratio statistic*, hereafter the LR statistic, equals

$$G^2(M_C \mid M_U) = 2 \ln L(M_U) - 2 \ln L(M_C)$$

Under very general conditions, if H_0 is true, then G^2 is asymptotically distributed as chi-square with degrees of freedom equal to the number of independent constraints. While the LR statistic can be used to compare any pair of nested models, there are two tests that are commonly computed by standard software and are often included in tables presenting the results of models estimated by ML.

The first test compares a given model to the constrained model in which all slope coefficients are equal to 0. This test is frequently referred to as the *likelihood ratio chi-square* or the *LR chi-square*. To define the test, let model M_β be the unconstrained model that includes an intercept, slope coefficients, and any other parameters in the model (e.g., σ in the LRM). Let M_α be the constrained model that excludes all regressors from the model (e.g., only parameters β_0 and σ would be included for the LRM). To test the hypothesis that all of the slope coefficients are simultaneously equal to 0, we use the test statistic:

$$G^2(M_\beta) = 2 \ln L(M_\beta) - 2 \ln L(M_\alpha) \qquad [4.6]$$

The simpler notation $G^2(M_\beta)$ replaces the more cumbersome $G^2(M_\alpha \mid M_\beta)$. If the null hypothesis that all slopes are 0 is true, then $G^2(M_\beta)$ is distributed as chi-square with degrees of freedom equal to the number of regressors.

The second test, known as the *scaled deviance* or simply the *deviance*, is used extensively within the framework known as the generalized linear model (McCullagh & Nelder, 1989, pp. 33–34). The deviance compares a given model to the *full model* M_F. The full model has one parameter for each observation, and can reproduce perfectly the observed data. Since the observed data are perfectly predicted, the likelihood of M_F is 1, and the log likelihood is 0. To test that M_F significantly improves the fit over M_β, the deviance is defined as

$$\begin{aligned} D(M_\beta) &= 2 \ln L(M_F) - 2 \ln L(M_\beta) \\ &= -2 \ln L(M_\beta) \\ &= G^2(M_\beta \mid M_F) \end{aligned}$$

Since the deviance is −2 times the log likelihood of the given model, its value can be computed readily from any program that provides the log likelihood of the model being estimated.

While $D(M_\beta)$ is sometimes reported as having a chi-square distribution, McCullagh (1986) shows that $D(M_\beta)$ has an asymptotic normal distribution as a consequence of the number of parameters in the full model increasing directly with the number of observations. McCullagh and Nelder (1989, pp. 120–122) suggest that when the data are *sparse* (i.e., when each combination of values of the independent variables occurs only once in the sample), $D(M_\beta)$ should not be used as a measure of fit in the model. See Hosmer and Lemeshow (1989, pp. 137–145) for further details.

$G^2(M_\beta)$ and $D(M_\beta)$ can be used to compare nested models. Consider the unconstrained model M_U and the constrained model M_C. If the values of the likelihood function are known, we could test the constraints on M_U with $G^2(M_C \mid M_U) = 2 \ln L(M_U) - 2 \ln L(M_C)$. This statistic could also be computed using the LR chi-squares:

$$\begin{aligned} G^2(M_U) &= 2 \ln L(M_U) - 2 \ln L(M_\alpha) \\ G^2(M_C) &= 2 \ln L(M_C) - 2 \ln L(M_\alpha) \end{aligned}$$

Since M_α is the same for both models,

$$\begin{aligned} G^2(M_C \mid M_U) &= G^2(M_U) - G^2(M_C) \\ &= 2 \ln L(M_U) - 2 \ln L(M_C) \end{aligned}$$

This is why $G^2(M_C \mid M_U)$ is often referred to as a *difference of chi-square test*. Similarly, the deviance can be used to compute the test. If

$$D(M_U) = -2 \ln L(M_U) \qquad \text{and} \qquad D(M_C) = -2 \ln L(M_C)$$

then

$$\begin{aligned} G^2(M_C \mid M_U) &= D(M_C) - D(M_U) \\ &= -2 \ln L(M_C) - -2 \ln L(M_U) \\ &= 2 \ln L(M_U) - 2 \ln L(M_C) \end{aligned}$$

Examples of the LR Test: Labor Force Participation

For the unconstrained model in Equation 4.5, the LR chi-square $G^2(M_U) = 124.48$ and the deviance $D(M_U) = 905.27$. These statistics are used for computing the following tests.

LR Test of a Single Coefficient. To test H_0: $\beta_1 = 0$, the model $M_{[K5]}$ is estimated, where the bracketed subscript indicates that *K5* is excluded from the unconstrained model. The LR chi-square and deviance for the constrained model are

$$G^2(M_{[K5]}) = 58.00 \qquad \text{and} \qquad D(M_{[K5]}) = 971.75$$

Then,

$$\begin{aligned} G^2(M_{[K5]} \mid M_U) &= G^2(M_U) - G^2(M_{[K5]}) = 66.48 \\ &= D(M_{[K5]}) - D(M_U) = 66.48 \end{aligned}$$

We conclude:

- The effect of having young children is significant at the .01 level ($LRX^2 = 66.5$, $df = 1$, $p < .01$).

Notice that I have used LRX^2 rather than G^2 in presenting the result. This makes it explicit that a likelihood ratio test is being reported.

LR Test of Multiple Coefficients. To test the hypothesis that the effects of the husband's and wife's education are simultaneously 0, H_0: $\beta_4 = \beta_5 = 0$, the model $M_{[WC,HC]}$ is estimated, resulting in

$$G^2(M_{[WC,HC]}) = 105.98 \qquad \text{and} \qquad D(M_{[WC,HC]}) = 923.76$$

The test statistic is

$$\begin{aligned} G^2(M_{[WC,HC]} \mid M_U) &= G^2(M_U) - G^2(M_{[WC,HC]}) = 18.50 \\ &= D(M_{[WC,HC]}) - D(M_U) = 18.50 \end{aligned}$$

We conclude:

- The hypothesis that the effects of the husband's and wife's education are simultaneously equal to 0 can be rejected at the .01 level ($LRX^2 = 18.5$, $df = 2$, $p < .01$).

LR Test That All Coefficients Are 0. $G^2(M_U) = G^2(M_\alpha \mid M_U)$ can be used to test the hypothesis that none of the regressors affects the probability of entering the labor force. Formally, H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$. We conclude:

- We can reject the hypothesis that all coefficients except the intercept are 0 at the .01 level ($LRX^2 = 124.5$, $df = 7$, $p < .01$).

While a Wald test could be used to test this hypothesis, the LR test is more commonly used.

TABLE 4.1 Comparing Results From the LR and Wald Tests

		LR Test		*Wald Test*	
Hypothesis	*df*	G^2	*p*	*W*	*p*
$\beta_1 = 0$	1	66.5	< 0.01	55.1	< 0.01
$\beta_4 = \beta_5 = 0$	2	18.5	< 0.01	17.7	< 0.01
All slopes = 0	7	124.5	< 0.01	95.0	< 0.01

4.1.4. Comparing the LR and Wald Tests

Even though the LR and Wald tests are asymptotically equivalent, in finite samples they give different answers, particularly for small samples. In general, it is unclear whether one test is to be preferred to the other. Rothenberg (1984) suggests that neither test is uniformly superior, while Hauck and Donner (1977) suggest that the Wald test is less powerful than the LR test. In practice, the choice of which test to use is often determined by convenience. While the LR test requires the estimation of two models, the computation of the test only involves subtraction. The Wald test only requires estimation of a single model, but the computation of the test involves matrix manipulations. Which test is more convenient depends on the software being used.

Table 4.1 compares the results of the LR and Wald tests for our example based on a sample of 753. For all hypotheses, the conclusions from both tests are the same. Note, however, that the values of the LR statistics are larger than the corresponding Wald statistics.

4.1.5. Computational Issues

There are two important computational considerations that must be taken into account when computing Wald and LR tests. If they are not, you run the risk of drawing the wrong conclusions from your tests.

Computing the LR Test

The LR test requires using the same sample for all models being compared. Since ML estimation excludes cases with missing data, it is common for the sample size to change when a variable has been excluded. For example, if x_1 has three missing observations that are not missing for any other variables, the usable sample increases by 3 when x_1 is excluded from the model. To ensure that the sample size does not change, you should construct a data set that excludes every observation that has

missing values for any of the variables used in any of the models being tested. Alternatively, missing values can be imputed using methods discussed in Little and Rubin (1987).

Computing the Wald Test

The matrix computations for the Wald test can accumulate appreciable rounding error if you do not use the full precision of the estimated coefficients and covariance matrix. Practically speaking, this means that you should use a program in which the estimates can be stored and then analyzed. Using the rounded values listed in the output can result in incorrect values for the test statistic.

4.2. Residuals and Influence

When assessing a model, it is useful to consider how well the model fits each case and how much influence each case has on the estimates of the parameters. *Residuals* measure the difference between the model's prediction for a given case and the observed value for that case, with observations that fit poorly thought of as *outliers*. *Influence* is the effect of an observation on estimates of the model's parameters or measures of fit. The analysis of residuals and influence is well developed for the LRM, and I assume that you have some familiarity with this material (see Fox, 1991, and Weisberg, 1980, Chapter 5, for good introductions). This section considers Pregibon's (1981) extensions of these methods to the BRM.

For a binary model, define $\pi_i = E(y_i \mid \mathbf{x}_i) = \Pr(y_i = 1 \mid \mathbf{x}_i)$. Since y is a binary variable, the deviations $y_i - \pi_i$ are heteroscedastic, with $\text{Var}(y_i \mid \mathbf{x}_i) = \pi_i(1 - \pi_i)$. This suggests the *Pearson residual*:

$$r_i = \frac{y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1 - \widehat{\pi}_i)}}$$

Large values of r_i suggest a failure of the model to fit a given observation. Pearson residuals can be used to construct a summary measure of fit, known as a *Pearson statistic*:

$$X^2 = \sum_{i=1}^{N} r_i^2$$

While X^2 is sometimes reported as having a chi-square distribution, McCullagh (1986) demonstrated that when the data are sparse (e.g., when there are continuous independent variables), X^2 has an asymptotic normal distribution with a mean and variance that are difficult to compute. McCullagh and Nelder (1989, pp. 112–122) recommended that X^2 not be used as an absolute measure of fit. Hosmer and Lemeshow (1989, pp. 140–145) propose an alternative test constructed by grouping data that can be used with sparse data.

While $\text{Var}(y_i - \pi_i) = \pi_i(1-\pi_i)$, $\text{Var}(y_i - \widehat{\pi}_i) \neq \widehat{\pi}_i(1-\widehat{\pi}_i)$. Consequently, the variance of r_i is not 1. To compute the variance of the estimated residuals, we need what is known as the *hat matrix*, so named because it transforms the observed y into $\widehat{y}$ in the LRM. For the BRM, Pregibon (1981) derived the hat matrix:

$$\mathbf{H} = \widehat{\mathbf{V}}\mathbf{X}(\mathbf{X}'\widehat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{V}}$$

where $\widehat{\mathbf{V}}$ is a diagonal matrix with $\sqrt{\widehat{\pi}_i(1 - \widehat{\pi}_i)}$ on the diagonal. Since only the diagonal of $\mathbf{H}$ is needed, we can use the computationally simpler formula:

$$h_{ii} = \widehat{\pi}_i(1 - \widehat{\pi}_i)\mathbf{x}_i \widehat{\text{Var}}(\widehat{\boldsymbol{\beta}})\mathbf{x}_i'$$

where $\mathbf{x}_i$ is a row vector with values of the independent variables for the ith observation and $\widehat{\text{Var}}(\widehat{\boldsymbol{\beta}})$ is the estimated covariance of the ML estimator $\widehat{\boldsymbol{\beta}}$. Using $1 - h_{ii}$ to estimate the variance of r_i, the *standardized Pearson residual* is

$$r_i^{\text{Std}} = \frac{r_i}{\sqrt{1 - h_{ii}}}$$

While r^{Std} is preferred to r, the two residuals are often similar in practice. An *index plot* of the standardized residuals against the observation number can be used to search for outliers. Figure 4.4 is an index plot of the standardized residual for the labor force data. Only half of the observations are shown in order to make the figure clearer. Two observations stand out as extreme and are marked with boxes. Observation 142 has a residual of 3.2; observation 512 has a residual of −2.7. Further analyses of these cases might reveal either incorrectly coded data or some inadequacy in the specification of the model. Cases with large positive or negative residuals should *not* simply be discarded from the analysis, but rather should be examined to determine why they were fit so poorly.

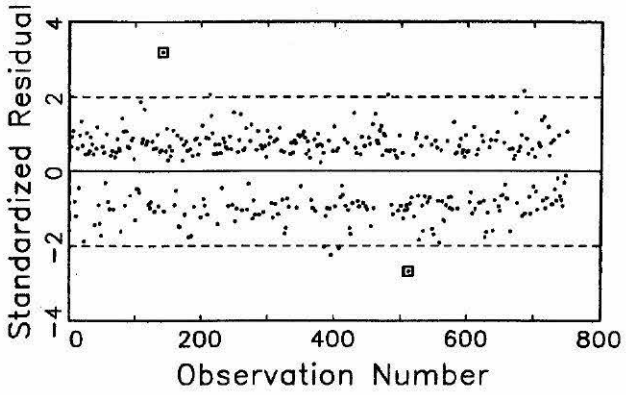

Figure 4.4. Index Plot of Standardized Pearson Residuals

While large residuals indicate that an observation is not fit well, they do not indicate whether an observation has a large influence on the estimated β's or the overall fit. For example, a large residual for the ith observation will not have a large influence on the estimates of β (i.e., removing that observation will not change the estimates very much) if $\mathbf{x}_i$ is near the center of the data. Being near the center of the data means that an observation's values for each independent variable are close to that variable's mean in the sample. On the other hand, extreme observations can influence the estimates, even when they do not have large residuals. A useful way to detect such observations, known as *high leverage* points, is to compute the change in $\widehat{\boldsymbol{\beta}}$ that occurs when the ith observation is deleted. Since it is computationally impractical to estimate the model N times, once with each observation removed, Pregibon (1981) derived an approximation that only requires estimating the model once. The expected change in $\widehat{\boldsymbol{\beta}}$ if the ith observation is removed is approximately equal to

$$\Delta_i\widehat{\boldsymbol{\beta}} = \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\mathbf{x}_i' \frac{y_i - \widehat{\pi}_i}{1 - h_{ii}}$$

The standardized change in β_k due to the deletion of $\mathbf{x}_i$, known as the *DFBETA*, equals

$$DFBETA_{ik} = \frac{\Delta_i\widehat{\beta}_k}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_k)}}$$

A large value of $DFBETA_{ik}$ indicates that the ith observation has a large influence on the estimate of β_k.

A second measure summarizes the effect of removing the ith observation on the entire vector $\widehat{\boldsymbol{\beta}}$, which is the counterpart to Cook's distance for the LRM:

$$C_i = (\Delta_i\widehat{\boldsymbol{\beta}})'\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})(\Delta_i\widehat{\boldsymbol{\beta}}) = \frac{r_i^2 h_{ii}}{(1 - h_{ii})^2}$$

Another measure of the impact of a single observation is the change in X^2 when the ith observation is removed:

$$\Delta_i X^2 = \frac{r_i^2}{1 - h_{ii}}$$

Figure 4.5 shows an index plot of C. Comparing this figure to Figure 4.4 illustrates the difference between an outlier and an influential observation. In both figures, observation 142 stands out. However, while observation 554 has a large residual, it has a C of only .06. Analysis of the $DFBETA_{ik}$'s for observation 142 would indicate which coefficients are being affected.

Methods for plotting residuals and outliers can be extended in many ways, including plots of different diagnostics against one another. Details of these plots are found in Landwehr et al. (1984) and Hosmer and Lemeshow (1989, pp. 149–170). While Lesaffre and Albert (1989) have proposed extensions of these diagnostics to the multinomial logit model, these extensions have not been added to standard software. Diagnostics for logit and probit are included in SAS and Stata.

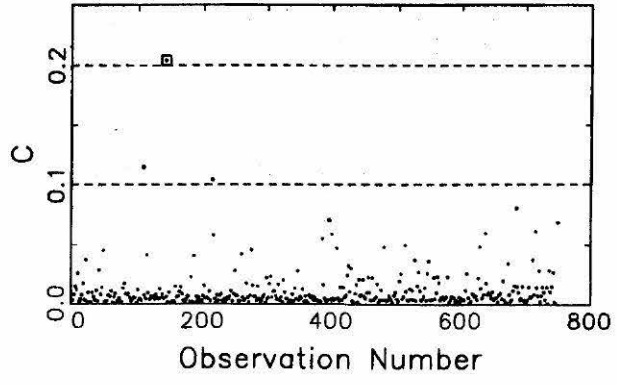

Figure 4.5. Index Plot of Cook's Influence Statistics

4.3. Scalar Measures of Fit

In addition to assessing the fit of each observation, it is sometimes useful to have a single number to summarize the overall goodness of fit of a model. Such a measure might aid in comparing competing models and, ultimately, in selecting a final model. Within a substantive area of research, measures of fit can provide a *rough* index of whether a model is adequate. For example, if prior models of labor force participation routinely have values of .4 for a given measure of fit, you would expect that new analyses with a different sample and perhaps with revised measures of the independent variables would result in a similar value for that measure of fit. Much larger or smaller values would suggest the need to reassess the changes made in the new study.

While the desirability of a scalar measure of fit is clear, in practice their use is problematic. First, I am unaware of convincing evidence that selecting a model that maximizes the value of a given measure of fit results in a model that is optimal in any sense other than the model having a larger value of that measure. While measures of fit provide some information, it is only partial information that must be assessed within the context of the theory motivating the analysis, past research, and the estimated parameters of the model being considered. Second, while in the LRM the coefficient of determination R^2 is the standard measure of fit, there is no clear choice for models with categorical outcomes. There have been numerous attempts to construct a counterpart to R^2 in the LRM, but no one measure is clearly superior and none has the advantages of a clear interpretation in terms of explained variation. Other measures have been constructed based on the ability of a model to predict the observed outcome. Finally, the Bayesian measures AIC and BIC, which are useful for comparing nonnested models, are increasingly popular. Overall, while I approach scalar measures of fit with some skepticism, their popularity and proliferation makes a review useful.

4.3.1. R^2 in the LRM

Many scalar measures of fit for models with CLDVs are constructed to approximate the coefficient of determination R^2 in the LRM. Most commonly, R^2 is defined as the proportion of the variation in y that can be explained by the x's in the model. However, R^2 can be defined in other ways, each of which produces an identical value for R^2 in the LRM. However, when these equivalent formulas are applied to models for CLDVs, they often produce different values and thus provide different measures of fit. [1]

Let the structural model be $y = \mathbf{x}\boldsymbol{\beta} + \varepsilon$, with K regressors, an intercept, and N observations. The expected value of y is $\hat{y} = \mathbf{x}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the OLS estimator. The coefficient of determination can be defined in each of the following ways. Derivations of these formulas can be found in Judge et al. (1985, pp. 29–31), Goldberger (1991, pp. 176–179), and Pindyck and Rubinfeld (1991, pp. 61, 76–78, 98–99).

The Percentage of Explained Variation Let $RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ be the sum of squared residuals, and let $TSS = \sum_{i=1}^{N}(y_i - \bar{y})^2$ be the total sum of squares. Then R^2 is the percentage of TSS explained by the x's:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \qquad [4.7]$$

The Ratio of $\widehat{\text{Var}}(y)$ *and* $\widehat{\text{Var}}(\hat{y})$ The ratio of the variances of $\hat{y}$ and y is another definition:

$$R^2 = \frac{\widehat{\text{Var}}(\hat{y})}{\widehat{\text{Var}}(y)} = \frac{\widehat{\text{Var}}(\hat{y})}{\widehat{\text{Var}}(\hat{y}) + \widehat{\text{Var}}(\hat{\varepsilon})} \qquad [4.8]$$

A Transformation of the Likelihood Ratio. If the errors are assumed to be normal, then R^2 can be written as

$$R^2 = 1 - \left[\frac{L(M_\alpha)}{L(M_\beta)}\right]^{2/N} \qquad [4.9]$$

where $L(M_\alpha)$ is the likelihood for the model with just the intercept, and $L(M_\beta)$ is the likelihood for the model including the regressors.

A Transformation of the F-Test. The hypothesis H_0: $\beta_1 = \cdots = \beta_K = 0$ can be tested using an F-test, with the test statistic F. R^2 can be written in terms of F as

$$R^2 = \frac{FK}{FK + (N - K - 1)}$$

where K is the number of independent variables.

[1] This is similar to the case in the LRM when there is no intercept. See Judge et al. (1985, pp. 30–31).

4.3.2. Pseudo-R^2's Based on R^2 in the LRM

Several pseudo-R^2's for models with CLDVs have been defined by analogy to the formula given in the last section. These formulas produce different values in models with categorical outcomes, and, consequently, are thought of as distinct measures.

The Percentage of Explained "Variation." For binary outcomes, Efron's (1978) pseudo-R^2 defines $\hat{y}$ as $\hat{\pi} = \widehat{\Pr}(y \mid \mathbf{x})$ and applies Equation 4.7:

$$R^2_{\text{Efron}} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

(*Show that in the case of a binary outcome,* $\sum_{i=1}^{N}(y_i - \bar{y})^2 = (n_0 n_1)/N$, *where* n_0 *is the number of* 0*'s and* n_1 *is the number of* 1*'s in the sample.*)

McFadden (1973) suggested a different analogy to explained variation in the LRM that can be applied to any model estimated with ML. This popular measure is also referred to as the "likelihood ratio index." In this measure, the log likelihood for model M_α without regressors is thought of as the total sum of squares, while the log likelihood of model M_β with regressors is thought of as the residual sum of squares. By analogy to Equation 4.7,

$$R^2_{\text{McF}} = 1 - \frac{\ln \widehat{L}(M_\beta)}{\ln \widehat{L}(M_\alpha)}$$

If model $M_\alpha = M_\beta$ (i.e., the slopes are all 0), R^2_{McF} equals 0, but R^2_{McF} can never exactly equal 1.

Like R^2 for the LRM, R^2_{McF} increases as new variables are added to the model. To compensate, Ben-Akiva and Lerman (1985, p. 167) suggest adjusting R^2_{McF} for the number of parameters in the model (just as the adjusted $\bar{R}^2$ in the LRM):

$$\bar{R}^2_{\text{McF}} = 1 - \frac{\ln \widehat{L}(M_\beta) - K}{\ln \widehat{L}(M_\alpha)}$$

$\bar{R}^2_{\text{McF}}$ will only increase if $\ln \widehat{L}(M_\beta)$ increases by more than 1 for each parameter added to the model.

Ben-Akiva and Lerman (1985, p. 167) discuss the logic behind and limitations of these measures. All else being equal, models with a larger value of the log likelihood are preferred, and $\bar{R}^2_{\text{McF}}$ provides a convenient

way to compare log likelihoods across different models. Unfortunately, there is no clear interpretation of values other than 0 and 1, nor is there any standard by which to judge if the value is "large enough."

The Ratio of $\text{Var}(y^*)$ *and* $\text{Var}(\hat{y}^*)$. For models defined in terms of a latent outcome according to $y^* = \mathbf{x}\boldsymbol{\beta} + \varepsilon$, McKelvey and Zavoina (1975, pp. 111–112) proposed a pseudo-R^2 by analogy to Equation 4.8:

$$R^2_{\text{M\&Z}} = \frac{\widehat{\text{Var}}(\hat{y}^*)}{\widehat{\text{Var}}(y^*)} = \frac{\widehat{\text{Var}}(\hat{y}^*)}{\widehat{\text{Var}}(\hat{y}^*) + \text{Var}(\varepsilon)}$$

This formula differs from that for the LRM in two respects. First, we are using the estimated variance of the latent variable y^* rather than the observed y. Second, the variance of ε is fixed by assumption, rather than being estimated. For the logit model, $\text{Var}(\varepsilon) = \pi^2/3$, and for the probit model, $\text{Var}(\varepsilon) = 1$. The variance of $\hat{y}^*$ can be computed as

$$\widehat{\text{Var}}(\hat{y}^*) = \hat{\boldsymbol{\beta}}' \widehat{\text{Var}}(\mathbf{x}) \hat{\boldsymbol{\beta}}$$

where $\widehat{\text{Var}}(\mathbf{x})$ is the estimated covariance matrix among the x's.

$R^2_{\text{M\&Z}}$ was suggested by McKelvey and Zavoina (1975, pp. 111–112) for ordinal outcomes, but can also be applied to binary and censored outcomes (Laitila, 1993). In simulation studies, Hagle and Mitchell (1992) and Windmeijer (1995) find that $R^2_{\text{M\&Z}}$ most closely approximates the R^2 obtained from regressions on the underlying latent variable.

A Transformation of the Likelihood Ratio. If we define M_α as the model with just the intercept, and M_β as the model with the regressors included, by analogy to Equation 4.9 a pseudo-R^2 can be defined as

$$R^2_{\text{ML}} = 1 - \left[\frac{L(M_\alpha)}{L(M_\beta)}\right]^{2/N} \qquad [4.10]$$

Maddala (1983, p. 39) shows that R^2_{ML} can be expressed as a transformation of the likelihood ratio chi-square $G^2 = -2\ln[L(M_\alpha)/L(M_\beta)]$:

$$R^2_{\text{ML}} = 1 - \exp(-G^2/N)$$

which illustrates that measures of fit such as R^2 and the various pseudo-R^2's are often closely related to tests of hypothesis. See Magee (1990) for other measures of fit based on the Wald and score tests.

TABLE 4.2 R^2-Type Measures of Fit for the Logit and LPM Models

	LPM		*Logit*	
Measure	M_1	M_2	M_1	M_2
$\ln L_\beta$	−478.086	−486.426	−452.633	−461.653
$\ln L_\alpha$	−539.410	−539.410	−514.873	−514.873
R^2_{Efron}	0.150	0.131	0.155	0.135
R^2_{McF}	0.114	0.098	0.121	0.103
$R^2_{\text{M\&Z}}$	0.150	0.131	0.217	0.182
R^2_{ML}	0.150	0.131	0.152	0.132
$R^2_{\text{C\&U}}$	0.197	0.172	0.205	0.177

NOTE: $N = 753$. $\ln L_\beta$ is the log likelihood for the full model; $\ln L_\alpha$ is the log likelihood for the model with no regressors; see the text for definitions of other measures.

As the fit of M_β approaches the fit of M_α [i.e., as $L(M_\beta) \to L(M_\alpha)$], R^2_{ML} approaches 0. However, Maddala (1983, pp. 39–40) shows that R^2_{ML} only reaches a maximum of $1 - L(M_\alpha)^{2/N}$. This led Cragg and Uhler (1970) to suggest the normed measure:

$$R^2_{\text{C\&U}} = \frac{R^2_{\text{ML}}}{\max R^2_{\text{ML}}} = \frac{1 - [L(M_\alpha)/L(M_\beta)]^{2/N}}{1 - L(M_\alpha)^{2/N}}$$

Since both R^2_{ML} and $R^2_{\text{C\&U}}$ are defined in terms of the likelihood function, they can be applied to any model estimated by ML.

Examples of Pseudo-R^2's: Labor Force Participation

To illustrate scalar measures of fit, consider two models. Model M_1 has the original specification of independent variables: *K5*, *K618*, *AGE*, *WC*, *HC*, *LWG*, and *INC*. Model M_2 adds a squared age term *AGE2* and drops the variables *K618*, *HC*, and *LWG*. The resulting measures of fit for the LPM and logit models are given in Table 4.2. Notice that for a given model many of the measures are identical for the LPM, but not for the logit model. You should try to reproduce these measures using the log likelihoods for the full and restricted models.

4.3.3. Pseudo-R^2's Using Observed Versus Predicted Values

Another approach to assessing goodness of fit in models with categorical outcomes is to compare the observed values to the predicted values. While I develop this idea for models with two outcomes, it can be easily generalized to models with J ordinal or nominal outcomes.

Let the observed y equal 0 or 1. The predicted probability that $y = 1$ is

$$\widehat{\pi}_i = \widehat{\Pr}(y = 1 \mid \mathbf{x}_i) = F(\mathbf{x}_i\widehat{\boldsymbol{\beta}}) \qquad [4.11]$$

where F is the cdf for the normal distribution for probit and for the logistic distribution for logit. Define the expected outcome $\widehat{y}$ as

$$\widehat{y}_i = \begin{cases} 0 & \text{if } \widehat{\pi}_i \le 0.5 \\ 1 & \text{if } \widehat{\pi}_i > 0.5 \end{cases}$$

which Cramer (1991, p. 90) calls the "maximum probability rule." This allows us to construct a table of observed and predicted values, such as Table 4.3, which is sometimes called a *classification table*.

The Count R^2. A simple and *seemingly* appealing measure based on the table of observed and expected counts is the proportion of correct predictions, which Maddala (1992, p. 334) refers to as the *count* R^2:

$$R^2_{\text{Count}} = \frac{1}{N}\sum_j n_{jj}$$

where the n_{jj}'s are the number of correct predictions for outcome j, which are located on the diagonal cells in Table 4.3.

The Adjusted Count R^2. The count R^2 can give the faulty impression that the model is predicting very well, when, in fact, it is not. In a binary model without knowledge about the independent variables, it is possible to correctly predict at least 50% of the cases by choosing the outcome category with the largest percentage of observed cases. For example, 57% of our sample were in the paid labor force. If we predict that all women are working, we would be correct 57% of the time. Accordingly,

TABLE 4.3 Classification Table of Observed and Predicted Outcomes for a Binary Response Model

Observed Outcome	*Predicted Outcome* $\widehat{y}=1$	$\widehat{y}=0$	*Row Total*
$y = 1$	n_{11} :: correct	n_{12} :: incorrect	n_{1+}
$y = 0$	n_{21} :: incorrect	n_{22} :: correct	n_{2+}
Column Total	n_{+1}	n_{+2}	N

R^2_{Count} needs to be adjusted to account for the largest row marginal. This can be done by

$$R^2_{\text{AdjCount}} = \frac{\sum_j n_{jj} - max_r(n_{r+})}{N - max_r(n_{r+})}$$

n_{r+} is the marginal for row r, so that $max_r(n_{r+})$ is the maximum row marginal (i.e., the number of cases in the outcome with the most observations). The *adjusted count* R^2 is the proportion of correct guesses beyond the number that would be correctly guessed by choosing the largest marginal, and can be interpreted as:

- Knowledge of the independent variables, compared to basing our prediction only on the marginal distributions, reduces the error in prediction by $100 \times R^2_{\text{AdjCount}}$%.

R^2_{AdjCount} is equal to Goodman and Kruskal's λ (Bishop et al. 1975, p. 388) applied to the classification table. Other measures of association could also be applied to the classification table (Menard, 1995 pp. 24–36).

Examples of Count Measures: Labor Force Participation

Table 4.4 shows the observed and predicted values from the logit model with independent variables: *K5*, *K618*, *AGE*, *WC*, *HC*, *LWG*, and *INC*. The row percentages indicate the percentage of a given outcome that were predicted to be either 1's or 0's. They show that the model is more effective at predicting 0's (80% are predicted correctly) than 1's (55% are predicted correctly). In this example, the count R^2 is

$$R^2_{\text{Count}} = \frac{180 + 342}{753} = .69$$

which can be compared to 57% of the cases that were observed as 1's. On the other hand, the adjusted R^2 is

$$R^2_{\text{AdjCount}} = \frac{(180 + 342) - 428}{753 - 428} = .29$$

shows that the models reduces the errors in prediction by 29%.

TABLE 4.4 Observed and Predicted Outcomes for the Logit Model of Labor Force Participation

	Predicted Outcome		
Observed Outcome	$\hat{y} = 0$	$\hat{y} = 1$	*Row Total*
$y = 0$	180	145	325
Row %	55.4	44.6	
$y = 1$	86	342	428
Row %	20.1	79.9	
Column Total	266	487	753
Row %	35.3	64.7	

4.3.4. Information Measures

A different approach to assessing the fit of a model and for comparing competing models is based on measures of information. Akaike's information criterion (AIC) is a well-known measure, while the Bayesian information criterion (BIC) is a measure that is gaining increasing popularity. For a general discussion of information-based measures, see Judge et al. (1985, pp. 870–875).

Akaike's Information Criterion (AIC)

Akaike's (1973) information criterion is defined as

$$\text{AIC} = \frac{-2 \ln \widehat{L}(M_\beta) + 2P}{N} \qquad [4.12]$$

where $\widehat{L}(M_\beta)$ is the likelihood of the model and P is the number of parameters in the model (e.g., $K + 1$ in the binary regression model where K is the number of regressors). While Akaike (1973) formally derives AIC through the comparison of a given model to a set of inferior alternative models, here I only provide a heuristic motivation for the reasonableness of the formula.

$\widehat{L}(M_\beta)$ indicates the likelihood of the data for the model, with larger values indicating a better fit. $-2 \ln \widehat{L}(M_\beta)$ ranges from 0 to $+\infty$ with smaller values indicating a better fit. As the number of parameters in the model becomes larger, $-2 \ln \widehat{L}(M_\beta)$ becomes smaller since more parameters make what is observed more likely. $2P$ is added to $-2 \ln \widehat{L}(M_\beta)$ as a penalty for increasing the number of parameters. Since the number of observations affects $-2 \ln \widehat{L}(M_\beta)$, we divide by N to obtain the

per observation contribution to the adjusted $-2\ln\widehat{L}(M_\beta)$. All else being equal, smaller values suggest a better fitting model.

AIC is often used to compare models across different samples or to compare nonnested models that cannot be compared with the LR test. All else being equal, the model with the smaller AIC is considered the better fitting model.

The Bayesian Information Criterion (BIC)

The Bayesian information criterion has been proposed by Raftery (1996, and the literature cited therein) as a measure to assess the overall fit of a model and to allow the comparison of both nested and nonnested models. This section summarizes Raftery (1996), which derives the formulas given below.

BIC is based on a Bayesian comparison of models. Consider models M_1 andM_2. The *posterior odds* of M_2 relative to M_1 equal

$$\frac{\Pr(M_2 \mid \text{Observed Data})}{\Pr(M_1 \mid \text{Observed Data})}$$

If the probability of M_2 given the observed data is greater than the probability of M_1 given the observed data, M_2 would be preferred. Under the assumption that the prior odds $Pr(M_2)/Pr(M_1)$ of the two models are 1 (i.e., we have no prior preference for one model over the other), the Bayes theorem can be used to show that the posterior odds equal the *Bayes factor*:

$$\frac{\Pr(\text{Observed Data} \mid M_2)}{\Pr(\text{Observed Data} \mid M_1)}$$

Model M_2 would be chosen if the probability of the observed data given that M_2 generated the data is greater than the probability of the observed data given M_1. Even if neither M_2 nor M_1 is the "true" model, the Bayes factor "is designed to choose the model that will, on average, give better out-of-sample predictions" (Raftery, 1996, p. 14).

The BIC statistic is a computationally convenient approximation to the Bayes factor. Given N observations, consider model M_k with deviance $D(M_k)$ comparing M_k to the saturated model M_S with df_k equal to the sample size minus the number of parameters in M_k.[2] The first

[2] In Section 4.1.3, I used the term "full model" to refer to what Raftery calls the saturated model.

BIC measure equals

$$\text{BIC}_k = D(M_k) - df_k \ln N \qquad [4.13]$$

Since BIC_S for the saturated model equals 0 (*Why must this be the case?*), the saturated model is preferred when $\text{BIC}_k > 0$. When $\text{BIC}_k < 0$, M_k is preferred with the more negative the BIC_k the better the fit.

A second version of BIC is based on the LR chi-square in Equation 4.6 with df_k' equal to the number of regressors (not parameters) in the model:

$$\text{BIC}_k' = -G^2(M_k) + df_k' \ln N \qquad [4.14]$$

If M_α is the null model without any regressors, then BIC_α' is 0. The null model is preferred when $\text{BIC}_k' > 0$, suggesting that M_k includes too many parameters or variables. When $\text{BIC}_k' < 0$, then M_k is preferred with the more negative the BIC_k' the better the fit. Basically, BIC_k' assesses whether M_k fits the data sufficiently well to justify the number of parameters that are used.

Either BIC_k or BIC_k' can be used to compare models, whether or not they are nested. Raftery (1996) shows that

$$2\ln\left[\frac{\Pr(\text{Observed Data} \mid M_2)}{\Pr(\text{Observed Data} \mid M_1)}\right] \approx \text{BIC}_1 - \text{BIC}_2 \qquad [4.15]$$

Thus, the difference in the BICs from two models indicates which model is more likely to have generated the observed data. Further, it can be shown that

$$\text{BIC}_1 - \text{BIC}_2 = \text{BIC}_1' - \text{BIC}_2'$$

so that the choice of which BIC measure to use is a matter of convenience.

Based on Equation 4.15, the model with the smaller BIC or BIC′ is preferred. How strong the preference is depends on the magnitude of the difference. Raftery, based on Jeffreys (1961), suggested guidelines for the strength of evidence favoring M_2 against M_1 based on a difference in BIC or BIC′. These are listed in Table 4.5. Since the model with the more negative BIC or BIC′ is preferred, if $\text{BIC}_1 - \text{BIC}_2 < 0$, then the first model is preferred. If $\text{BIC}_1 - \text{BIC}_2 > 0$, then the second model is preferred.

TABLE 4.5 Strength of Evidence Based on the Absolute Value of the Difference in BIC or BIC′

Absolute Difference	*Evidence*
0–2	Weak
2–6	Positive
6–10	Strong
> 10	Very Strong

Finally, to see the link between BIC and other measures of fit, consider the formula that Raftery (1996, p. 19) provides for computing BIC′ in the LRM:

$$\text{BIC}'_k = N \ln(1 - R_k^2) + df_k \ln N$$

This convenient computational formula for BIC′ in the LRM can also be used for models with CLDVs by replacing R_k^2 by R_{ML}^2 from Equation 4.10.

Example of Information Measures: Labor Force Participation

To illustrate the AIC and BIC measures, the logit model M_1 with the original specification of independent variables: *K5*, *K618*, *AGE*, *WC*, *HC*, *LWG*, and *INC*; and M_2 which adds a squared age term *AGE2* and drops the variables *K618*, *HC*, and *LWG* were estimated. Table 4.6 contains the test statistics, along with the components that are used to compute them. Since many programs do not compute the AIC and BIC, it is important to verify that you can obtain the listed statistics using the formula in Equations 4.12 through 4.14.

Based on the values of AIC, BIC, and BIC′, model M_1 is favored by all measures. Using the difference in BIC,

$$\text{BIC}_1 - \text{BIC}_2 = -4{,}029.66 - -4{,}024.87 = -4.79$$
$$\text{BIC}'_1 - \text{BIC}'_2 = - \quad 78.11 - - \quad 73.32 = -4.79$$

According to Table 4.5, the evidence favoring M_1 over M_2 is positive but not strong.

4.4. Conclusions

The methods for hypothesis testing are quite general and can be used with all models considered in this book. Pregibon's methods for detecting

TABLE 4.6 AIC and BIC for the Logit Model

Measure	M_1	M_2
$\ln L_g$	−452.633	−461.653
$\ln L_a$	−514.873	−514.873
G^2	124.481	106.441
D	905.266	923.306
df	745	747
df'	7	5
P	8	6
AIC	1.223	1.242
BIC	−4029.663	−4024.871
BIC′	−78.112	−73.321

NOTE: $\ln L_g$ is the log likelihood for the full model; $\ln L_a$ is the log likelihood for the model with no regressors. $N = 753$.

outliers and influential observations apply only to models with binary outcomes. While some of the scalar measures of goodness of fit are only appropriate for models with binary outcomes, others apply with minor adjustments to any model estimated with ML.

4.5. Bibliographic Notes

The tests presented in this chapter have a long history. R. A. Fisher introduced the LR test in the 1920s, and A. Wald proposed the Wald test in the 1940s. Further details on these tests can be found in most econometrics texts. Godfrey (1988, pp. 8–20) and Cramer (1986, pp. 30–42) contain thorough discussions of the foundations of these tests. Buse (1982) provides an informative geometric interpretation. Maddala (1992, pp. 118–124) presents an interesting discussion within the context of the linear regression model. Regression diagnostics for the binary response model were developed by Pregibon (1981). Amemiya (1981) and Windmeijer (1995) have reviews of measures of fit. Hosmer and Lemeshow (1989, Chapter 5) provide further details on diagnostics and tests of fit. The AIC was proposed by Akaike (1973). The BIC has been advocated by Raftery in a series of papers summarized in Raftery (1996). His work developed from Schwarz (1978) and Jeffreys (1961). See Judge et al. (1985, pp. 870–875) for a discussion of these and related measures.