CHAPTER 5

Regression Methods and Correlation

We run into some problems when we try to apply sample size determination methods to regression models, as will be seen in this chapter. The types of problems encountered vary with the type of model used and the number of regression variables in the model.

There are many important applications of regression over a wide variety of fields, however, so it is highly desirable to try to overcome these problems. For example, many colleges and universities have used regression prediction equations to predict a prospective student's college grade point average (GPA) if the student were admitted, and to make an admittance decision based partly on that result. Several researchers have investigated the minimum sample size for regression models developed for this purpose and this is discussed in Section 5.1.2.1.

5.1 LINEAR REGRESSION

The general form for a multiple regression model is typically written

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \epsilon \tag{5.1}$$

with Y the dependent variable, $X_1, X_2, \ldots, X_m$ the independent variables (also called the predictors or regressors), and ϵ the error term, with the error terms assumed to be independent and to have a normal distribution with a mean of zero and a common variance of σ_ϵ^2.

The simplest approach to determining sample size is to adopt the Draper and Smith (1998) rule-of-thumb and use at least 10 observations for each predictor in a linear regression model. This recommendation had also been made by Halinksi and Feldt (1970) and Miller and Kunce (1973). That isn't possible in many types of

Sample Size Determination and Power, First Edition. Thomas P. Ryan.

applications, however, especially in chemometric applications where the number of predictors will often exceed the number of observations. So determination of sample size for regression problems will often be necessary, with the sample size that is obtained perhaps compared with the sample size using the Draper and Smith (1998) rule-of-thumb.

5.1.1 Simple Linear Regression

The simple linear regression model is

$$Y = \beta_0 + \beta_1 X + \epsilon$$

That is, there is just a single independent variable, X.

Work on determining sample size to estimate β_1 when both X and Y are random (the usual case in practice but not in textbooks) dates at least from Thigpen (1987).

Unfortunately, if we take an analytical approach to sample size determination, we can run into problems, depending on how we proceed, even when there is only a single predictor. Adcock (1997) explained that it is necessary not only to specify the nonzero value of β_1 that constitutes the alternative hypothesis in simple linear regression, but also to supply estimates of σ_ϵ^2 and the variance of the predictor values. Of course, the latter can be estimated reasonably well if the predictor values are to be fixed, but it may be very difficult to estimate if the values are not fixed. For random X, the same approach might be used as was discussed in Section 2.1; that is, use the range of likely values divided by 6 in place of s_x.

We can avoid those problems, however, if we do not focus on testing β_1 but instead specify the value of r^2 (the square of the Pearson correlation between X and Y) that would be considered acceptable and that we want to detect. This would be for a test of $\rho = 0$ as the null hypothesis, with ρ designating the population correlation between X and Y. That is, the test would be outside the realm of regression but would apply to the regression model. This is perfectly reasonable because if a simple linear regression model is written in "correlation form," the slope coefficient is the correlation between X and Y. That is, the model would be written as $Y - \bar{Y} = \beta^*(X - \bar{X}) + \epsilon$, with β^* then being the correlation coefficient.

Another problem, not addressed by Adcock (1997), is that it is not sufficient just to reject H_0: $\beta_1 = 0$, as the model could still have poor predictive value. We can also avoid this problem by focusing on r^2 and using sample size determination software that has this option.

Despite this, undoubtedly many users will want to focus on β_1 when it has an easy and important physical interpretation. Therefore, we will address how to proceed when that is the focal point. As explained by Ryan (2009a, p. 20), if we

adapt the work of Wetz (1964) to the test of H_0: $\beta_1 = 0$, we would conclude that the model is useful only if the value of the t-statistic for testing the hypothesis is at least twice the critical value. We can actually use this rule-of-thumb to circumvent the problem of having to specify values of σ_ϵ^2 and the variance (or the standard deviation) of the predictor values, although in doing so we do run into some other problems.

This is illustrated in the following example.

■ EXAMPLE 5.1

Consider the use of Lenth's applet, which requires that the user enter values for s_x, the standard deviation of the X-values, s_e, the standard deviation of the residuals (the observed values minus the values from the fitted model), and the "detectable beta," which is defined as "the clinically meaningful value of the regression coefficient that you want to be able to detect." The input requirements for PASS are essentially the same, except that the user must enter the standard deviation of the residuals rather than the error standard deviation, which is essentially just a semantic difference. (There are many other options for sample size determination software for simple linear regression, including the user-written command `sampsi_reg` for Stata.)

When a t-test is used to test H_0: $\beta_1 = 0$, the t-statistic is

$$t = \frac{\hat{\beta}_1}{s_e/\sqrt{S_{xx}}}$$

with $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Since $\sqrt{n-1}\, s_x = \sqrt{S_{xx}}$, it follows that we may write the t-statistic as

$$t = \frac{\hat{\beta}_1}{s_e/(\sqrt{n-1}s_x)} = \left(\frac{s_x}{s_e}\right)\hat{\beta}_1\sqrt{n-1}$$

If we apply the adaptation of the Wetz (1964) rule-of-thumb, we would want to have $t = (s_x/s_e)\hat{\beta}_1\sqrt{n-1} \geq 2t_{\alpha/2,n-2}$ for a two-sided test. Since the right side will slightly exceed 4 for most sample sizes, we might enter $s_x = s_e = \hat{\beta}_1 = 1$ into an applet or software. This would be a reasonable starting point because both power and the value of the t-statistic depend only on the ratio, s_x/s_e, not on the individual values. Furthermore, these are the default values that the user of Lenth's applet encounters. The objective would be to see if the sample size that is produced is at least 18, since the left side of the last inequality would then be at least $\sqrt{17} = 4.12$, and thus be at least slightly greater than 4, as desired. Lenth's applet produces $n = 11$ if power is to be at least .80, with power being .8385 for that sample size.

Similarly, when PASS is used, entering "1" for "Standard Deviation of Residuals" and for "Standard Deviation of *Xs*," with slope of 0 specified for B0 (the null hypothesis) and 1 for Bl (the alternative hypothesis), power = .80 and $\alpha = .05$ for a two-tailed test, the software gives $n = 11$ and power = .8385, in agreement with Lenth's applet.

If a sample of that size were taken resulting in $\hat{\beta}_1 = 1$ and $s_x = s_e$, the value of the t-statistic would be $\sqrt{10} = 3.16$, which is far short of $2t_{\alpha/2,n-2} = 2t_{.025,9} = 2(2.262) = 4.524$. Thus, although the power is .8385 for detecting a value of β_1 that is 3.16 times the standard error of $\hat{\beta}_1$, the Wetz (1964) criterion is not met because 3.16 is less than 4.524. Consequently, it would be a good idea to increase the sample size and ignore a target value of .80 for the power, as that is just an arbitrary value anyway. It is necessary to use $n \geq 19$ in order for the Wetz criterion to be met; using $n = 19$ produces a power of .9839. This satisfies the Draper–Smith rule-of-thumb (as did $n = 11$) and may also satisfy the Wetz criterion, depending on the sample values.

This ad hoc procedure seems to be a reasonable way to overcome the problems posed by the input requirements for determining sample size when an applet or software is used, especially since quantities such as s_e depend on how well the model explains the dependent variable, and a value for s_x will generally be hard to determine if the values of the regressor are not preselected. (Regressor values are random in most applications.) Furthermore, the focal point should be determining the sample size that will enable the user to determine if the model is useful, not being able to detect a (perhaps) arbitrary value of β_1 for an arbitrary power value of .80. Therefore, even if there is a theoretical basis for an experimenter to believe that β_1 should be equal to a certain value in a particular application, it would still be desirable to use the ad hoc approach given in this section and compare the sample sizes that are obtained using each approach, using the result from the ad hoc approach as a benchmark. Of course, the ad hoc approach does rely on inputted values that almost certainly won't be the true values, but that will always be the case whenever sample sizes are determined in any type of application.

We can contrast the results for this example with the result obtained using the Soper applet for linear regression (`http://www.danielsoper.com/statcalc/calc01.aspx`). With that applet, the user enters the number of predictors, value of α, minimum acceptable power, and effect size, and the applet returns the sample size. For example, using $\alpha = .05$ with simple linear regression, a minimum acceptable power of .80, and an effect size of 2.33, the applet gives a sample size of 7.

It is easier to relate to an R^2 value in regression than to relate to an effect size, which is more appropriate for experimental design applications and doesn't have an obvious interpretation in regression problems. The relationship between the two is given by $R^2 = f^2/(f^2 + 1)$, with f^2 the usual notation for effect size.

Thus, $f^2 = R^2/(1 - R^2)$, which doesn't have a meaningful physical interpretation since it is the proportion of the variation in the dependent variable that is explained by the model divided by the proportion that is not explained by the model.

It was determined for this example that $n = 19$ should be used. Since $t_{.025,17} = 2.10982$, $2t_{.025,17} = 4.21963$. We can use the latter to solve for R^2 since the relationship between the value of the t-statistic and the value of R^2 in simple linear regression is given by $R^2 = t^2/(n - 2 + t^2)$. [See, for example, Ryan (2009a, p. 20).] Performing the computation produces $R^2 = .5516$, which leads to $f^2 = 1.04737$.

Since $n = 19$ produces power = .9839 using Lenth's applet and the same for PASS, we would hope that entering $\alpha = .05$, power = .9839, and effect size = 1.04737 in Soper's applet would produce a sample size of 19, which indeed is what happens.

What is the practical value of all of this? It is useful to consider certain rules-of-thumb such as those given by Draper and Smith (1981) and Wetz (1964), but of course we won't know if the latter is satisfied until we take a sample, whose desirable size we are trying to determine! If the regressor values will be selected, then s_x can be computed, or at least estimated, and a detectable value for β_1 should not be difficult to determine. With Lenth's applet and PASS, however, it is also necessary to specify a value for s_e, in addition to the desired power, although in PASS a value for r_{xy}, the correlation between X and Y, can be used in lieu of s_e. The latter would be a good choice if the user has a good idea of what the correlation might be, and should certainly be easier to specify than s_e, which doesn't have the intuitive appeal of a correlation coefficient.

Soper's applet is easier to use for linear regression as the detectable effect size can be determined without difficulty if the applet user has a minimum acceptable R^2 value in mind, as it is then just a matter of solving for the effect size from the value of the smallest acceptable R^2 value. Then it is just a matter of specifying the minimum acceptable power and seeing if the sample size that results is acceptable when other criteria (e.g., sampling cost) are considered. Note that this does not incorporate the Wetz (1964) criterion, but specifying a minimum acceptable R^2 value essentially serves as a replacement for that criterion since, as indicated previously, R^2 is a function of t^2. Given the selected sample size, a user might then take a sample of that size and see if the Wetz criterion is met. If so, then the regression equation should be at least reasonably useful.

Soper's applet uses an iterative procedure to arrive at the sample size, with the procedure explained at `http://www.danielsoper.com/statcalc/calc01.aspx`.

It is important to note that no method of determining sample size will ensure that the resultant regression equation will be useful because that depends on the relationship between X and Y exhibited by the data. ■

5.1.2 Multiple Linear Regression

The multiple linear regression case is far more difficult than simple linear regression and practically intractable because the values of multiple linear regression parameters generally don't have any meaning when the predictor values are random (Ryan, 2009a, p. 168). If the parameter values don't have any meaning, then a confidence interval on a parameter value similarly would not have any meaning, and it then logically follows that determining sample size so as to be able to detect regression parameters with specific parameter values also seems highly questionable. This is apparently not well understood, however, as Maxwell (2000) and Kelley and Maxwell (2003) discussed determining sample size in such a way as to give a confidence interval for each β_i that is sufficiently narrow. [*Note*: Other authors of regression books have not taken such a strong position regarding the interpretation of regression parameters and coefficients, but Cook and Weisberg (1999) stated: "In some problems, this simple description of a regression coefficient cannot be used, particularly if changing one term while holding others fixed doesn't make sense .... In a study of the effects of economic policies on quality of life, for example, changing one term like the prime interest rate may necessarily cause a change in other possible terms like the unemployment rate. In situations like these, interpretation of coefficients can be difficult."]

Recognizing the complexities involved, Adcock (1997) stated: "What to do in the multiple-regression case remains a topic for further study."

More recently, Dattalo (2008, p. 29) gave effect sizes of small, medium, and large for R^2, following the general idea of Cohen (1988) of designating effects of 0.02, 0.15, and 0.35 as small, medium, and large, respectively. Since $R^2 = f^2/(f^2 + 1)$, with f^2 denoting the effect size, the corresponding values of R^2 are .02, .13, and .26, for small, medium, and large, respectively. Obviously, f^2 and R^2 will be equal, to two decimal places, when f^2 is close to zero. Lenth (2001) appropriately criticized the use of the small, medium, and large designations, which clearly doesn't make any sense for regression models, and the same can be said for other applications. Indeed, Cohen (1988), who used effect size throughout his book, admitted (p. 413) that there wasn't a strong need for it in regression in stating "... the need to think in terms of f^2 is reduced, and with it, the need to rely on conventional operational definitions of 'small,' 'medium,' and 'large' values for f^2. We nevertheless offer such conventions for the frame of reference that they provide, and for use in power surveys and other methodological investigations." My advice is to simply forget about f^2 altogether.

An acceptable value of R^2 depends on the particular application and, as a rough rule-of-thumb, R^2 generally varies inversely with sample size and also increases as regressors are added to a model, but for most applications, $R^2 \geq .90$ is generally considered desirable, although much smaller values of R^2 are viewed as important in some applications. Thus, the threshold values of R^2 that result from the arbitrary designation of effect sizes given by Cohen (1988) will not lead to acceptable R^2 values for most applications.

In contrast to the values given by Dattalo (2008), we could proceed as follows. Since $R^2 = SSR/SST$, with SSR the sum of squares due to regression and SST the total sum of squares, the F-statistic for testing the significance of a regression model can be written

$$F = \frac{\frac{R^2}{k}}{\frac{1 - R^2}{n - k - 1}} \tag{5.2}$$

for a multiple linear regression model with k regressors. The reader is asked to verify this result in Exercise 4.1, which can easily be done using the fact that F is (more customarily) defined as

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n - k - 1}}$$

Since the Wetz (1964) criterion was defined in terms of the F-statistic for the regression model, with the criterion being that the computed F-value should be at least four times the critical value, we could use that to obtain a benchmark for a "good" R^2 value by solving Eq. (4.2) for R^2 as a function of F, and then applying the Wetz criterion.

Solving for R^2 produces

$$R^2 = \frac{kF}{n - k - 1 + kF} \tag{5.3}$$

The Wetz (1964) criterion could then be applied to produce

$$R^2_{\text{desired}} = \frac{k4F_{.05,k,n-k-1}}{n - k - 1 + k4F_{.05,k,n-k-1}} \geq a \tag{5.4}$$

with a a number such as .7 or .8. Of course, this lower bound on desirable R^2 values depends on the sample size through the $F_{.05,k,n-k-1}$ tabular value, but this might be used to solve for n for fixed k and for a benchmark R^2 value. Alternatively, following the original discussion, Eq. (5.4) might be used to identify desirable R^2 values for given values of n and k. For example, with $k = 3$ and $n = 30$ (just meeting the 10:1 rule-of-thumb), we obtain $R^2_{\text{desired}} \geq .58$. This is a reasonable lower bound because a lower value than this would not be acceptable in the vast majority of applications involving three regressors.

Alternatively, we could solve for n for fixed k and a lower bound on R^2, such as, say, .80. If we set R^2_{desired} in Eq. (5.4) at .80 and use $k = 3$, we obtain

$n = 4.0 + 3F_{.05,3,n-4}$. We obtain $n = 14.8$ from inspection of an F-table, which would round to $n = 15$. This does not come close to satisfying the 10:1 rule-of-thumb, although that guideline will not be met in many applications.

We can see from Eq. (5.4) that R^2 will be large, using the Wetz (1964) criterion, whenever the Wetz criterion is met and $n - k - 1$ is small, approaching 1 as $n - k - 1$ approaches 0. (Of course, there is an exact fit whenever the sample size equals the number of model parameters, which is the same as saying that $n - k - 1 = 0$.) Making that quantity small, however, would be an artificial way of trying to produce a desirable R^2 value. This relates to the discussion in Draper and Smith (1981) regarding how R^2 can be inflated when the number of observations is not much larger than the number of distinct sets of regressor values.

We can't make a direct comparison of the results obtained using the Wetz (1964) criterion and the results obtained using Soper's applet because the former does not involve power. Nevertheless, it is of interest to see some results using Soper's applet for $k = 3$. Using the latter, if we want to detect $R^2 \geq .80$ (so that the minimum detectable effect is 4.0) and set power nominally at .80 with $\alpha = .05$, the applet gives $n = 8$. With such a sample size, the error variance would be estimated with only four degrees of freedom. Furthermore, the n:k ratio is poor. Power would have to be set at .9998 or higher in order to obtain $n = 15$, which was the value obtained based on the Wetz (1964) criterion.

Soper's applet gives a small value of n when a large value of R^2 is used (to solve for f^2) because the objective is to *detect* a large effect when the null hypothesis is that the effect size is zero. Wetz's criterion, on the other hand, is not used for sample size determination, but could be used for that purpose in a roundabout way, as was done earlier in this section. Since power is not involved in Wetz's criterion, we might suspect that the sample sizes obtained indirectly through the use of that criterion may correspond to the results obtained using Soper's applet if the power selected for the latter is virtually 1.0, as was the case in the example just given.

Soper's applet uses the noncentral F-distribution, whereas Wetz's criterion uses the central F-distribution, thus making a direct analytical comparison somewhat difficult. A question could be raised as to whether or not such a comparison is even appropriate since the Wetz criterion was not intended to be used to solve for sample size, although it could be used to do so indirectly.

Nevertheless, the following computations may be illuminating. If we let $a = .7$ in Eq. (5.4) and solve for n, we obtain $n \leq k + 1 + (12k/7)F_{.05,k,n-k-1}$. Of course, this is not a simple expression for n because the third term on the right side of the inequality is a function of n, so we can't obtain a closed-form expression. It is simple to solve for n by using an F-table, however, and it is more convenient to write the inequality as $n - k - 1 \leq (12k/7)F_{.05,k,n-k-1}$, since the left side is the denominator degrees of freedom for the F-statistic. By multiplying the entries in each column of the table by $(12k/7)$, we can then look

Table 5.1 Sample Sizes for Models with $k = 1$–6 Predictors Under Certain Sample Size Guidelines

	Wetz		Power = .8		
k	($R^2 > .70$, Power = .8)	($R^2 > .70$)	"Large" Effect (0.35, $R^2 = .26$)	"Medium" Effect (0.15, $R^2 = .13$)	"Small" Effect (0.02, $R^2 = .02$)
1	6	11	25	55	387
2	9	16	31	68	476
3	10	19	36	77	539
4	12	24	40	85	590
5	13	28	43	92	635
6	14	32	46	98	674

for the largest value of $n - k - 1$ that is less than the corresponding value of $(12k/7)F_{.05,k,n-k-1}$. For example, when $k = 2$, $n - k - 1 = 13$ is slightly less than $(24k/7)F_{.05,2,13} = 13.063$. Therefore, $n = 16$ is used.

Alternatively, we might let $a = .8$, reasoning that R^2 should be at least .8 for sample sizes typically used if the model has at least two or three predictors and the correlations between them are not large. This would result in smaller sample sizes selected than are obtained when $a = 0.7$, however, because the multiplier of $F_{.05,k,n-k-1}$ is then k instead of $12k/7$.

Consider Table 5.1, which gives the required sample sizes obtained through indirect use of the Wetz criterion (in column 2). The other columns give the sample sizes obtained by using the Power and Precision software, with the R^2 values entered, not the effect sizes, as there is no option to enter the latter. There are some differences, especially for the "small" effect, when those R^2 values are converted to effect sizes with four decimal places and those values are entered in Soper's applet. The differences are small, usually 1, except at the small effect, where the differences are about 2% of the Power and Precision values. (Similarly, differences exist if four decimal place R^2 values are inputted in Power and Precision corresponding to each of the three exact effect sizes.)

Since power is not part of the Wetz (1964) criterion, we might guess that the power associated with each of the Wetz sample sizes would be virtually 1. It can be shown using Soper's applet that the power values do range from .99 for $k = 1$ to .99999+ for $k = 6$, with the "+" indicating that the power is greater than .99999 but the Soper applet allows only five decimal places.

The values in the last three columns, using Cohen's three effect sizes, are shown simply for comparison, as the R^2 values are too low to be of practical value and the sample size values in the last column are absolutely nonsensical, as regression datasets are generally much smaller than that.

Although we might view the sample sizes obtained from indirect use of the Wetz criterion as being somewhat smaller than desirable, there are many well-known regression datasets with approximately these combinations of k and n. For

example, the famous stack loss dataset of Brownlee (1960) has $k = 3$ and $n = 21$ and the Gorman and Toman (1966) dataset had $k = 6$ and $n = 31$, differing only slightly from the corresponding Wetz (k, n) combinations. Both of these datasets have been analyzed extensively in the literature, and are used for illustration in Ryan (2009a).

Of course, we should bear in mind that Wetz (1964) did not propose sample size guidelines; I am simply obtaining guidelines by indirect use of the Wetz criterion. The sample sizes given are simply those that would result in R^2 being slightly in excess of .70 if the Wetz criterion is met when a set of regression data is analyzed.

One point that is sometimes made (`http://www.listserv.uga.edu/cgi-bin/wa?A2=ind0107&L=spssx-l&P=29839`) is that the requisite sample size depends not just on the number of predictors but also on the correlations between them. The logic behind such thinking is undoubtedly that the amount of information provided by the predictors is lessened when there is at least moderate correlations between them, necessitating a larger sample than would be required if the predictor correlations were small. Predictor correlations are not incorporated into the Power and Precision software nor into Soper's applet. They are also not built into PASS or nQuery, and the latter cannot be used to determine sample size for multiple regression with a general number of predictors. If suitable software were available, these correlations would have to be specified a priori, which would generally be very difficult or impossible without the availability of prior data. Nevertheless, sample size determination methods that incorporate correlations between the predictors have been proposed by Hsieh, Bloch, and Larsen (1998).

When PASS is used for determining sample size in multiple linear regression, the user specifies the number of predictors that will be included in the regression model without being tested and the R^2 value that should result from the inclusion of those predictors. (Zero is one of the options for the number of such predictors; correlations between predictors is not an input option.) Then the user specifies the number of predictors that are to be tested and the increase in R^2 that one wishes to detect. If the incremental R^2 is large, the necessary sample size will be small since the null hypothesis is that the incremental R^2 is zero. For example, if no predictors are to be automatically included and four predictors are to be tested, with desired power $= .90$ $\alpha = .05$, with $R^2 = .70$ anticipated for those four predictors, PASS gives $n = 13$. This is too small a sample size for a model with four predictors, however, but a small value of n should be expected because of the large difference between 0 and .70. If we substitute in Eq. (5.3), we obtain $R^2 = .66$. Even though this is close to the .70 used in the PASS sample size determination, there is no reason why they should be the same since a value for power is being specified for PASS but power is not involved in Eq. (5.3), nor in its development.

If the predictor values are fixed, as when an experimental design is used, the correlations between the predictor values would then be either zero or very

small numbers, so software that did not incorporate predictor intercorrelations as input would not be a problem. Notice that even with predictor intercorrelations, which opens up the possibility that all of the predictors would be used if variable selection methods were applied to the dataset, the PASS input requirement for multiple linear regression is much simpler than is the case for simple linear regression.

Other software for determining sample size relative to R^2 includes a program described by Mendoza and Stafford (2001).

Krishnamoorthy and Xia (2008) considered the case of sample size determination for testing a nonzero null hypothesis value of R^2.

If there is interest simply in determining sample size so that R^2 can be estimated with a suitably small error of estimation, this was addressed by Algina and Olejnik (2000).

5.1.2.1 Application: Predicting College Freshman Grade Point Average

An important sample size question in the use of regression models is: "How large should the sample size be for developing a regression model to predict what a prospective student's college freshman grade point average would be if the student were admitted?" Many colleges and universities use such regression models and base their admittance decisions partly on the predictions obtained using such models. Of course, the worth of such models depends on whether or not all, or at least almost all, of the important variables are in the model. That is an issue of variable selection and will not be discussed here.

Similar to other sample size determination problems, the parameters of the selected regression model must be estimated, and the larger the sample size, the smaller the variance of those estimators and the smaller will be the variance of the predictions. Sawyer (1983) concluded that total group predictions based on 70 or more students works as well as predictions obtained using larger samples, whereas separate prediction equations for males and females should be satisfactory if based on as few as 50 students.

5.2 LOGISTIC REGRESSION

The logistic regression model is usually written

$$\pi\left(X_1, X_2, \ldots X_k\right) = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k\right)} \tag{5.5}$$

for the general case of k regressors, $X_1, X_2, \ldots, X_k$, with $\pi(X_1, X_2, \ldots, X_k) = P(Y = 1|X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$, with the dependent variable, Y,

being binary. Alternatively, since $1 - \pi = 1/[1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k]$, $\pi/(1 - \pi) = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$, so

$$\log[\pi/(1 - \pi)] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

thus writing the right-hand side in the form of a linear model.

Hosmer and Lemeshow (2000, p. 339) stated: "There has been surprisingly little work on sample size for logistic regression. The available methods to address sample size selection have been implemented in just a few specialty software packages." Not much has changed since they made that statement.

Throughout this book there has been discussion of the use of pilot studies to obtain parameter estimates that can be used as software input to determine sample sizes, with internal pilot studies having some advantages over external pilot studies. Flack and Eudey (1993) illustrated the use of an external pilot study, a dental study involving 19 people. They cautioned, however, that the parameter estimates in logistic regression are, at least when maximum likelihood is used, large-sample approximations, but only a small sample (of 19) was used in this study to obtain the parameter estimates. Consequently, they suggested that a sensitivity analysis be performed by varying the parameter estimates and observing the resultant changes in the associated sample size values. Muller and Pasour (1997) examined the bias that results when only significant tests from a prior study result in power being calculated for the main study.

One point that cannot be overemphasized is that a continuous response variable should *not* be dichotomized, as doing so discards information and this results in a loss of power for a fixed sample size, or the need for a larger sample size for a fixed power. See, for example, Taylor, West, and Aiken (2006).

As discussed by Ryan (2009a, Chapter 9), there are more complexities inherent in multiple logistic regression than in multiple linear regression, and this also applies to sample size determination. As in linear regression, we will first consider the case of simple logistic regression.

5.2.1 Simple Logistic Regression

If we wanted to develop an approach to sample size determination analogous to what was done in Section 5.1.1 for simple linear regression, we would need a measure of model adequacy. Software utilizes the Wald test, which is analogous to the *t*-test in linear regression. Although the Wald test has received some criticism in the literature, we will use it here since it is used frequently in practice.

In linear regression, ordinary least squares is usually the method of estimation, so sample size determination software tacitly assumes the use of that method. In logistic regression, maximum likelihood is generally used, but will often not be the best estimation method (see the discussion in Ryan, 2009b). In particular, the maximum likelihood solution will not exist when complete separation exists. An example of this is when there is one continuous covariate and all of the values of

the covariate at $Y = 1$ are greater than all of the values of the covariate at $Y = 0$. Near separation exists when there is only slight overlap of the two sets of values and when that happens the maximum likelihood estimates will tend to blow up, as will the standard errors of their estimators. Because of this problem and other problems with maximum likelihood estimators in logistic regression, various alternative approaches have been proposed but they will not all be discussed in this chapter.

Nevertheless, we will initially assume the use of maximum likelihood and will start with the sample size determination methods that have been proposed. Whittemore (1981) was probably the first to propose a method for determining sample size in logistic regression and provided tables of required sample sizes. Those tables were for power values of .90, .95, and .99, but not .80. Graphs were also given to aid in sample size selection. The recommended approach was for small response probabilities only, however (say, $\pi < .1$), and when that is the case, the data will generally be sparse (i.e, there will be only a small percentage of $Y = 1$ values), necessitating the use of some method other than maximum likelihood.

Whittemore (1981) did assume the use of maximum likelihood, however, and showed that the proposed method is quite sensitive to the distribution of the corresponding covariates. Very large sample sizes can result from application of the method, with an example discussing possible sample sizes of 15,425 and 17,584. Even though this is an era in which larger amounts of data are being handled, those would be extremely large datasets.

Another sample size determination method was proposed by Self, Mauritsen, and Ohara (1992), but their method can be used only with discrete predictors, not continuous predictors. [Their method was extended to the class of generalized linear models by Shieh (2000).] See also Bull (1993), who considered the special case of a three-level predictor, including a three-level ordinal predictor, and Alam, Rao, and Cheng (2010), who proposed a variation of the Whittemore (1981) method and also critiqued the Hsieh et al. (1998) methods, pointing out that they are very sensitive to the choice of β_0. Therefore, they suggested that a pilot study be performed so that a reasonable value for β_0 might be selected.

Dattalo (2008, p. 32), for example, pointed out that sample size determination in simple logistic regression can be determined for a normally distributed covariate by using the sample size formula for a two-sample *t*-test, as previously stated by Hsieh et al. (1998) and Vaeth and Skovlund (2004). This follows from the fact that for a normally distributed covariate, the log odds value β_1 is zero if and only if the group means for the two response categories, assuming equal variances, are the same. Since Y has two values, 0 and 1, there are thus "two groups" of values, as in a two-sample *t*-test. The general idea with this approach is to avoid the considerable complexities inherent in sample size computations for logistic regression, as indicated here in Section 5.2 and also mentioned by Vaeth and Skovlund (2004). They stated that their method is exact when applied to linear regression, and approximate when applied to logistic regression and Cox regression

because it relies on asymptotic theory. Of course, we should generally be concerned about the latter, especially when the use of software leads to a sample size that is not very large. Simulations were performed that showed the adequacy of their method, although there were some problems with small samples and highly skewed predictor distributions.

Similarly, sample size could be determined by focusing on the odds ratio, which is defined as $(\pi/(1-\pi))$, with $\pi = P(Y = 1|X = x)$, as indicated previously, and $1 - \pi = P(Y = 0|X = x)$. That is, the odds ratio implicitly incorporates the two groups, as in the two-sample *t*-test. The applet of Demidenko (2007) solves for sample size in simple logistic regression for a binary covariate (only) with user-inputted values of the proportion of $X = 1$ values and $P(Y = 1|X = 0)$, as illustrated in the following example.

■ EXAMPLE 5.2

Using the applet of E. Demidenko (`http://www.dartmouth.edu/~eugened/power-samplesize.php`), if we specify power = .80, $\alpha = .05$, 50% of the *X*-values are "1", and $P(Y = 1|X = 0) = .50$, which would correspond to a binary covariate having no predictive ability at all, and specify a "detectable" odds ratio of 3, the applet gives a sample size of $n = 121$. Of course, an odds ratio of 3 would imply $\pi = .75$ and $1 - \pi = .25$, which would be a reasonable odds ratio. Of course, a specified odds ratio larger than 3 would require a smaller sample size, with $n = 84$ for an odds ratio of 4, and a larger sample size needed for a smaller odds ratio. For example, $n = 278$ is required for an odds ratio of 2.

Although the applet does not provide a graph of sample size plotted against detectable odds ratio as an option, such a graph can be produced by using a reasonable number of data points. Doing so produces the graph in Figure 5.1.

The computed sample sizes are based on the assumed use of the Wald test, which corresponds to the *t*-test in linear regression, as stated previously. Despite that correspondence, the Wald test has received some criticism [see, e.g., Hosmer and Lemeshow (2000, p. 17)]. The applet also assumes the use of maximum likelihood, which, as stated previously, would be a poor choice of estimation method for certain data configurations, as well as small sample sizes. Nevertheless, the applet results can be used as at least a rough general guideline, and Demidenko (2008) defends the use of the Wald test. ■

5.2.1.1 Normally Distributed Covariate

The software Power and Precision can also be used to obtain sample sizes for logistic regression models, as can other software including SiZ and PASS. [Broadly, the `powercal` command in Stata can be used to compute sample size and power for generalized linear models, a class that includes logistic regression models. See Newson (2004) for details.]

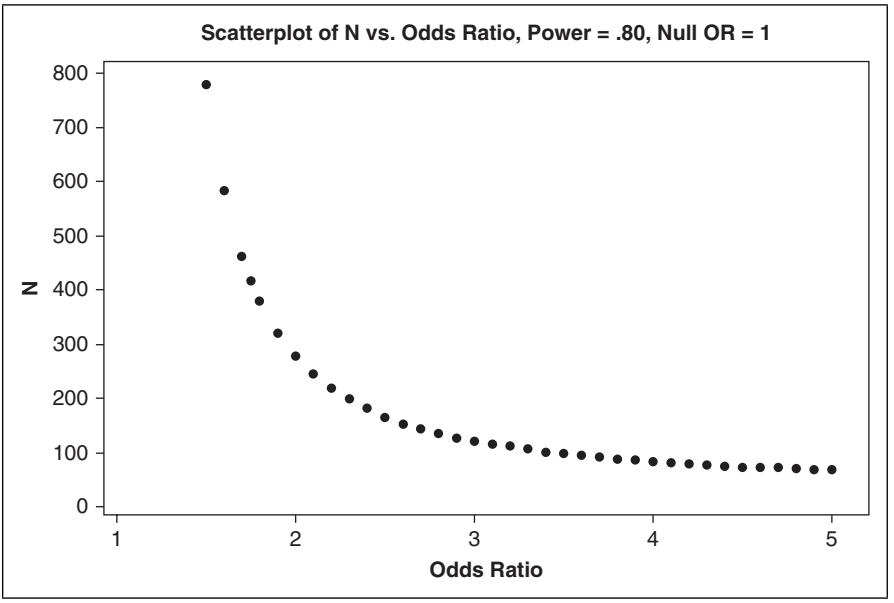

Figure 5.1 Relationship between sample size and detectable odds ratio; power = .80; null hypothesis: equal proportion of $x = 0$ and $x = 1$; $P(Y = 1|X = 0) = .5$.

For the assumption that X has a normal distribution, PASS uses the following expression, given by Hsieh et al. (1998):

$$n = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{\pi^*\left(1 - \pi^*\right) B^2} \tag{5.6}$$

with $z_{\alpha/2}$ and z_{β} as previously defined, $\pi^* = P(Y = 1)$ at the mean of X, μ_x, and $B = \sigma_x \beta_1$, with β_1 determined from $P(Y = 1)$ at $\mu_x + \sigma_x$.

For example, assume that H_0: $\beta_1 = 0$ is tested using a two-sided test with $\alpha = .05$, $\sigma = 5$, and power is to be .90. To obtain B expressed as a multiple of the standard deviation of X, we could proceed as follows. For example, if we wish to detect $P(Y = 1) = .65$ at $\mu_x + \sigma_x$, this assumption implies an assumption of a specific value of β_1, which in turn leads to the corresponding value of B. This can be explained as follows. It is well known [see, e.g., Ryan (2009a, p. 314)] that $P(Y = 1) = .5$ occurs at $X = -\beta_0/\beta_1$. [This can be verified by substitution in Eq. (5.5).] Let that value of $X = \mu_x$ so that $X = \mu_x + \sigma_x = -\beta_0/\beta_1 + 5$. Then from Eq. (5.5) we have

$$\pi(X) = \frac{\exp\left(\beta_0 + \beta_1 X\right)}{1 + \exp\left(\beta_0 + \beta_1 X\right)}$$

so that

$$.65 = \frac{\exp\left[\beta_0 + \beta_1\left(-\beta_0/\beta_1 + 5\right)\right]}{1 + \exp\left[\beta_0 + \beta_1\left(-\beta_0/\beta_1 + 5\right)\right]}$$

and thus

$$.65 = \frac{\exp\left(5\beta_1\right)}{1 + \exp\left(5\beta_1\right)}$$

Solving this equation for β_1, we obtain $\beta_1 = [\log(.65/.35)]/5 = 0.123808$. Since, as indicated, $B = \sigma_x\beta_1$ in Eq. (5.6), $B = 5(0.123808) = 0.619038$. Then, using Eq. (5.6), we obtain

$$\begin{aligned} n &= \frac{\left(z_{\alpha/2} + z_\beta\right)^2}{\pi^*\left(1 - \pi^*\right)B^2} \\ &= \frac{(1.95996 + 1.28155)^2}{(0.5)(0.5)(0.619038)^2} \\ &= 109.678 \end{aligned}$$

so that $n = 110$ would be used. This is the sample size given by SiZ, whereas PASS gives $n = 109$.

Some discrepancies between software can be expected because of the use of different sample size formulas. For example, nQuery uses a different formula [given as Eq. (4) in Hsieh et al. (1998)], so when nQuery is used, the sample size is given as $n = 138$, which obviously differs considerably from the sample sizes given by SiZ and PASS. The formula used by nQuery is

$$n = \frac{\left[z_{\alpha/2} + z_\beta \exp\left(-\beta^2/4\right)\right]^2}{\pi_1\beta^2}\left[1 + 2\pi_1\Delta\right]$$

with

$$\Delta = \frac{\left[1 + \left(1 + \beta^2\right)\exp\left(-5\beta^2/4\right)\right]}{1 + \exp\left(-\beta^2/4\right)}$$

Here β^2 is the square of the log odds value, $\beta = \log[\pi_2(1 - \pi_1)/(\pi_1(1 - \pi_2)]$, with π_1 and π_2 denoting the value of π at $X = \mu_x$ and $X = \mu_x + \sigma_x$, respectively.

Hsieh et al. (1998) suggested that this formula not be used when the odds ratio $\pi_2(1 - \pi_1)/\pi_1(1 - \pi_2)$ is at least 3, since it will then overestimate the required sample size. Here it is 1.86 and the fact that the sample size for this example was

much larger than the sample sizes obtained using PASS and SiZ was noted. This raises the question of whether the formula should be used at all.

Note that B is not part of the input with either PASS or Power and Precision, for example, as with Power and Precision the user would enter $P(Y = 1)$ at a value for X other than μ_x, in addition to entering $P(Y = 1)$ at μ_x, which we will later designate as $P_1(\mu_x)$. With PASS, the user would enter the value for $P(Y = 1)$ at μ_x and at $\mu_x + \sigma_x$, or enter the first of these two and the odds ratio.

Regarding these software input requirements, the following statement by Novikov, Fund, and Freeman (2010) is worth noting, with P_1 used in place of π and $E(X) = \mu_x$: "As, in our experience, the investigator is unlikely to recognize the difference between P_1 and $P1(E(X))$ and will generally supply the more natural P_1, we explore the impact that this will have on the sample sizes calculated by these methods." The sidebar explanation that is provided in PASS indicates that when both probabilities are entered, the first one is the probability at μ_x and the second one corresponds to $\mu_x + \sigma_x$. This is the same input requirement for SiZ and similarly the user of nQuery has the same two input options, with this clearly indicated by both software packages. What is being inputted into Power and Precision is also clearly indicated but, as with the other software, this does place a burden on the user with the user perhaps not being clear on what is required, as suggested by Novikov et al. (2010), or simply overlooking the explanations. In referring to nQuery, PASS, and Power and Precision, Novikov et al. (2010) also stated: "The above-mentioned packages use different algorithms for the calculation, and in certain circumstances the resulting estimates can be quite different, especially when the hypothesized effect is large."

The example that they gave to illustrate that is as follows. With $\alpha = .05$ for a two-sided test, power $= .80$, $P_1(\mu_x) = 0.5$ and the user wants to detect an odds ratio as large as 3.49, PASS gives the solution as $n = 20$, whereas the solution from Power and Precision is $n = 44$, as can be verified by a user who has both software packages, although the user would have to first compute $P(Y = 1|X = \mu_x + \sigma_x)$ for input in Power and Precision, which is .778. Obviously, there is a huge difference in the two sample sizes!

Accordingly, Novikov et al. (2010) proposed a method that utilized the sample size formula for a two-sample t-test with unequal variances and unequal sample sizes given by Schouten (1999) and given previously in Section 3.7. (As indicated previously, the two-sample idea comes from the fact that there are two groups of observations for Y: 0 and 1.) The method proposed by Novikov et al. (2010) is somewhat involved, however, and requires three preliminary steps before Schouten's formula is used, one of which involves solving an equation numerically. They provided a SAS program for all of the necessary computations.

Novikov et al. (2010) performed simulation studies to examine the accuracy of their proposed method and compared it to the methods of Hsieh et al. (1998) and Demidenko (2007). They also examined the effect of the user specifying

$P1(E(X))$ instead of P_1, which is the overall probability that $Y = 1$. Power was determined using the Wald test. Novikov et al. (2010) found that all three methods performed well for detecting small effects but that their proposed method is superior for detecting large effects. They also found that the method given by Hsieh (1989), which is an extension of the formula given by Whittemore (1981) and is used by the software nQuery, is based on the sampling distribution of the log of the odds ratio and overestimates the sample size, especially for a large response probability. As Novikov et al. (2010) stated, this result is not surprising since the method given by Whittemore (1981) was intended to be used for a small response probability, as stated previously. (Overestimation was evident in Example 5.2.)

5.2.1.2 Binary Covariate

Hsieh et al. (1998) gave the following sample size formula for a single binary covariate. In their notation it is

$$n = \frac{\left\{z_{\alpha/2}\left[P\left(1 - P/B\right)\right]^{1/2} + z_{\beta}\left[P1\left(1 - P1\right) + P2\left(1 - P2\right)\left(1 - B\right)/B\right]^{1/2}\right\}^2}{\left(P1 - P2\right)^2\left(1 - B\right)} \tag{5.7}$$

Here B denotes the proportion of the sample with $X = 1$; $P1$ and $P2$ are the event rates at $X = 0$ and $X = 1$, respectively; and the overall event rate, P, is given by $P = (1 - B)P1 + BP2$.

Hosmer and Lemeshow (2000, p. 339) stated that, for a single binary covariate, sample size determination is equivalent to determining the sample size for testing the equality of two proportions, and stated that the sample size formula for the latter can be used. That requires some clarification, however, because the $\bar{p}$ that is embedded in Eq. (5.6) is just a simple average of two proportions under the assumption of equal sample sizes, whereas the P in Eq. (5.7) is a weighted average since the proportion of responses at $X = 0$ and $X = 1$ probably won't be the same. Hosmer and Lemeshow (2000) gave an example using Eq. (5.6) and compared the sample size obtained using that formula with the sample size obtained using the Whittemore (1981) approach, with the latter being much larger than the sample size obtained using Eq. (5.6).

When software is used, the user must specify whether the covariate has a normal distribution or is binary. For example, in PASS the user selects "logistic regression" and then indicates whether the covariate is normally distributed or binary. In Power and Precision, the user first selects "Logistic" and then chooses the option for either "Logistic regression, one continuous predictor" (normality is not mentioned but that is apparently assumed) or "Logistic regression, one categorical predictor (two levels)." Not all software provides these options, however, as the simple logistic regression routines in nQuery and SiZ have the capability only for a normal covariate.

5.2.2 Multiple Logistic Regression

The simplified approach discussed by Hsieh et al. (1998) and Vaeth and Skovlund (2004) for simple logistic regression can, according to these authors, be applied to multiple logistic regression by using variance inflation factors (VIFs). As explained by Hsieh et al. (1998), Whittemore (1981) showed that for a set of normally distributed covariates, $\text{Var}(\widehat{\beta}_i)$ can be approximated by inflating the variance when only the corresponding predictor is used in the model. The inflation is accomplished by dividing that variance by $(1 - R_i^2)$, with R_i^2 denoting the square of the correlation between X_i and the other covariates in the model. [This is the VIF that is used in linear regression; see Ryan (2009a, p. 170).]

The necessary sample size is inflated in the same way. That is, the sample size necessary for testing $\beta_i = 0$ is inflated by dividing it by $(1 - R_i^2)$. Hsieh et al. (1998) reported that this inflation factor also "seems to work well for binary covariates."

Hosmer and Lemeshow (2000, p. 346) were critical of this ad hoc approach, however, stating: "We think that the sample size suggested by equation (8.49) may be unnecessarily large but could be the starting point for a more in depth sample size analysis using pilot data [to] do some model fitting." They distinguish between the sample size needed to detect a particular nonzero value of a model parameter with a desired power and the number of observations needed to fit a selected model.

There can be problems interpreting coefficients in logistic regression models due to multicollinearity, just as happens with linear regression models when multicollinearity is present, so determining sample size from hypothesis tests of regression parameters is not necessarily a good idea. It would be better to use the increase in R^2 when the covariate that is being tested is added as a criterion in testing a covariate and its parameter, but such an approach would have its own problems because there are multiple R^2 statistics that have been proposed for logistic regression and no one statistic stands out as being clearly superior to the other ones. See Menard (2000) and Sharma (2006). This would hinder any attempt to focus attention on determining sample size so as to try to ensure that the logistic regression model will be of value in terms of explaining the variability in the response variable. Therefore, developing a sample size determination approach that parallels the development for linear regression using the Wetz criterion (as in Sections 5.1 and 5.1.2) would be even more challenging than for linear regression and will not be attempted here. (Since such an approach has apparently not been given in the literature, it is also not available in software since software methods generally follow methods given in the literature, and generally lag the publication of those methods by several years.)

The concern that Hosmer and Lemeshow (2000) expressed over the sample size inflation factor given by Hsieh et al. (1998) is indirectly supported by the results of Peduzzi, Concato, Kemper, Holford, and Feinstein (1996), who found

that at least 10 events per covariate are needed to avoid problems of bias in the regression coefficients in both directions and similarly both underestimation and overestimation of variances of the estimators. For general models in which the number of parameters is greater than the number of covariates, they concluded that there should be at least 10 events *per parameter*.

Peduzzi et al. (1996) are indirectly stating that the necessary increase in the sample size should be (at least roughly) a linear function of the number of covariates or the number of parameters, if the two differ.

An online sample size rule-of-thumb that was motivated by the results of Peduzzi et al. (1996) is $n = 10k/p$ (see `http://www.medcalc.org/manual/logistic_regression.php`), with $k =$ the number of covariates and $p =$ the smaller of the anticipated proportion of events and nonevents in the population. Long (1997, p. 54) stated in referring to maximum likelihood (ML): "It is risky to use ML with samples smaller than 100." So a person who felt that way would probably want to round $10k/p$ up to 100, if necessary, although $10k/p$ might exceed 100 if the model has more than a few covariates or the dataset is sparse (i.e., unbalanced), so that p is small.

This would be a simple rule-of-thumb for practitioners to follow, whereas the Hsieh et al. (1998) adjustment depends on an adjustment factor that may be unknown, and will be only monotonically related to an increase in the number of covariates. More specifically, when a variable selection algorithm is applied to linear regression or logistic regression, a point of diminishing returns is reached (as measured, say, by R^2) in terms of addition of covariates. This is often, but not always, due to multicollinearity if the data are observational. If a designed experiment is used, however, R_i^2 may be increasing very little, if at all, when covariates are added. Then the Hsieh et al. (1998) adjustment factor would fail badly, according to the results of Peduzzi et al. (1996). The Hsieh et al. (1998) adjustment factor is also flawed in another way as it is inappropriate to use a Pearson correlation coefficient or multiple correlation coefficient to measure the correlation between continuous covariates and binary covariates, a mixture that occurs quite frequently with logistic regression data.

This is useful information because software capability for multiple logistic regression is limited. PASS and nQuery use the Hsieh et al. (1998) inflation factor as the user specifies the value of R_i^2 and there is no restriction on the number of covariates. The user of SiZ specifies $\sqrt{R_i^2}$. Power and Precision will handle only two continuous predictors and the user specifies the correlation between them.

So what should a user do? The problem with trying to use the results of Peduzzi et al. (1996) is that the user needs to know approximately what percentage of events and nonevents to expect for the covariates of interest. Software could then be used to arrive at a sample size, with the user facing a dilemma if the percentage multiplied times the sample size from the software output is not at least 10 times the number of covariates or 10 times the number of parameters in the tentative

model, whichever is appropriate. If that condition is not met, it might be safer to use a sample size that is large enough that the expected number of events meets the ≥ 10 rule-of-thumb.

Logistic regression is a generalized linear model. Self and Mauritsen (1988) considered sample size determination for such models, with logistic regression models considered as a special case. Lindsey (1997) gave a simple general formula for exact calculations of sample size for any member of the linear exponential family. This includes generalized linear models with a known or fixed dispersion parameter. Kodell, Lensing, Landes, Kumar, and Hauer-Jensen (2010) considered a specific application involving radiation measures and considered sample size deteremination when a logistic model is used as well as other generalized linear models.

5.2.2.1 Measurement Error

For the case of two continuous covariates, with one being an exposure variable that is subject to measurement error, an applet is given at `http://biostat.hitchcock.org/MeasurementError/Analytics/PowerCalculationsforLogisticRegression.asp` and is based on the work of Tosteson, Buzas, Demidenko, and Karagas (2003), which allows the incorporation of covariate measurement error, with the user specifying (estimating) the correlation between the observed measurement and the true measurement, as well as the correlation between the exposure variable and the other covariate.

This applet will not compute sample size, however, but will give the power for a selected sample size as one point on a graph that shows the power for various values of the odds ratio for a one standard deviation increase in the exposure variable. The graph actually has two lines/curves—one for measurement error and the other for no measurement error, as shown in Figure 5.2. Of course, such graphs are always useful because the inputted values will almost certainly not be the true values.

5.2.3 Polytomous Logistic Regression

The response variable in logistic regression can have more than two categories. When this is the case, it is called *polytomous logistic regression*. This is used much less frequently than binary logistic regression. Readers are referred to Section 8.1 of Hosmer and Lemeshow (2000) for information on polytomous logistic regression. Software for sample size determination does not generally include capability for polytomous logistic regression. Another problem is that polytomous logistic regression can be undermined by near separation just as in binary logistic regression, and the maximum likelihood estimators will not exist when there is compete separation. (How the latter can occur in polytomous logistic regression should be apparent from the discussion in Section 5.2.1, as it is just a matter of extending that discussion to more than two levels.)

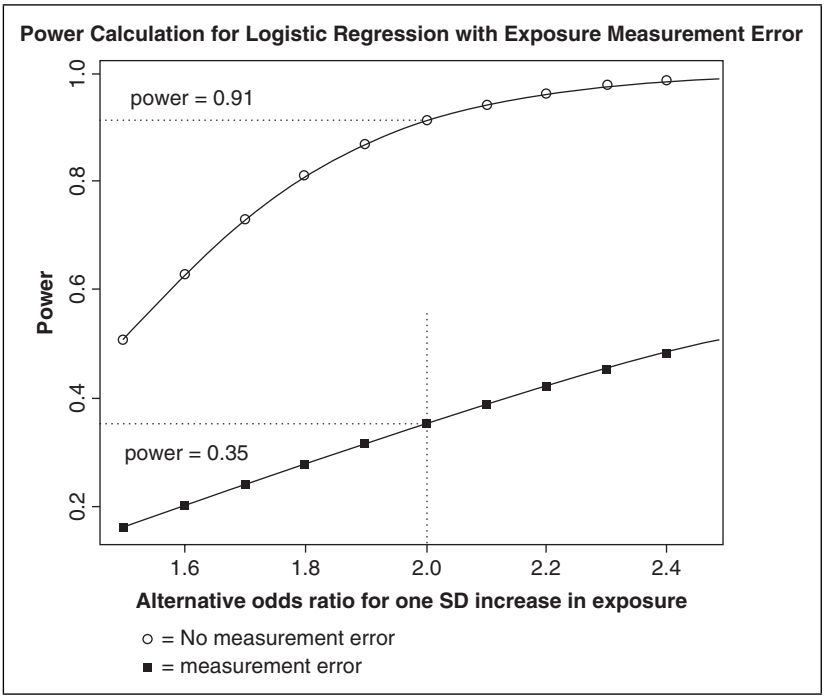

Figure 5.2 Power calculation for logistic regression with exposure measurement error. Correlation between true and observed exposure = 0.5. Odds ratio for covariate = 1.5. Prevalence with no exposure = 0.5. Correlation between exposure and covariate = 0. Sample size = 100. Significance level = 0.05.

5.2.4 Ordinal Logistic Regression

In both binary and polytomous logistic regression, the levels of the dependent variable are nominal. This is appropriate if the response variable is, say, gender or race, but not if, for example, ranks of people are involved, such as executive titles or army ranks.

Assume that the response variable has k ordered categories and there is a treatment group and a control group of subjects for each category. The formula for sample size, as given by Whitehead (1993) [see also Walters (2004) and Simon (2008)], is

$$n = \frac{6\left(Z_{\alpha/2} + Z_{\beta}\right)^2}{\log(OR)^2\left[1 - \sum_{i=1}^{k} \bar{\pi}_i^3\right]}$$

with OR denoting the odds ratio given in Section 5.2.1 and $\bar{\pi}_i$ being the average of the proportion of patients in the treatment and control groups for the ith category.

Sample size determination for ordinal logistic regression models has also been addressed by Vaughan and Guzy (2002), who presented an algorithm for exploring subsets of data from previous studies to enable a study to be designed with a desired power. Its use with SAS Software was illustrated.

See Section 8.2 of Hosmer and Lemeshow (2000) for information on ordinal logistic regression models, as they cover three different types of such models.

5.2.5 Exact Logistic Regression

One potential problem with all of the sample size methods for logistc regression that have been given in the literature is that they all assume the use of maximum likelihood (ML) as the method of estimation. As shown by King and Ryan (2002), the use of ML can result in very poor parameter estimates under certain conditions. There are a few alternatives to maximum likelihood in logistic regression, one of which is *exact logistic regression* (Cox, 1970, pp. 44–48). See also Mehta and Patel (1995). Quoting from the SAS Software documentation at `http://support.sas.com/rnd/app/da/new/daexactlogistic.html`: "Exact logistic regression has become an important analytical technique, especially in the pharmaceutical industry, since the usual asymptotic methods for analyzing small, skewed, or sparse data sets are unreliable." Ammann (2004) showed in a medical application of ML logistic regression, which had a small sample, that the exact p-value for testing for the inclusion of one of the predictors was .046, whereas the ML p-value was .231, thus giving totally different pictures regarding the worth of that predictor. Ammann (2004) stated that it would not have been possible to increase the number of patients without considerable effort. Thus, an estimation method that is not encumbered by small-sample bias should have been used. Such bias is well known (see, e.g., Firth, 1993), as the ML estimators are only asymptotically unbiased. So there could be major problems when there are small samples, as in the example cited by Ammann (2004).

Another alternative is a method due to Firth (1993), which might be the best overall method to use, although that has not yet been determined. Firth's method is penalized maximum likelihood, which reduces the amount of bias in the maximum likelihood estimator by adding a small amount of bias into the score function.

In addition to SAS Software, exact logistic regression is available in other software, including LogXact, Egret, and Stata. Sample size determination for exact logistic regression is apparently not available with any software, however.

5.3 COX REGRESSION

Cox regression, also known as proportional hazards regression, is used very extensively in survival analysis. Ryan and Woodall (2005) indicated that Cox (1972) was the second most frequently cited statistics paper and it now has well over 20,000 citations.

The Cox model can be written

$$h_Y(t) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \tag{5.8}$$

where $X_1, X_2, \ldots, X_k$ denote a set of predictor variables, $h_Y(t)$ denotes the hazard function of Y at time t, $h_0(t)$ is an arbitrary baseline hazard function computed, at time t, by possibly using zero as the value of each predictor variable, and the β_i are parameters to be estimated. [The hazard function is the probability of failure (i.e., death in survival analysis) at time t divided by the probability of survival until time t.]

The parameters are estimated using Cox regression and their interpretation is similar to the interpretation of the parameters in a multiple logistic regression model.

As in logistic regression, there are different, equivalent forms of the Cox regression model, with a linear model representation resulting from dividing both sides of Eq. (5.8) by $h_0(t)$ and then taking the log of each side of the equation, so as to give

$$\log\left(h_Y(t)/h_0(t)\right) = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{5.9}$$

Of course, this is very similar to what is done in logistic regression to produce a linear model. Another similarity is that maximum likelihood is used to estimate the parameters, with iterations performed until the log(likelihood) appears to have converged. Unfortunately, the use of maximum likelihood in Cox regression can sometimes cause the same type of problems that can exist with logistic regression. Specifically, the parameter estimates can either blow up or fail to exist. This was demonstrated in Ryan (2009b). Thus, there will be conditions under which an alternative to Cox regression should be used, and one such alternative is the method due to Firth (1993), which was mentioned in Section 5.2.5 and which was developed for the class of generalized linear models. That class includes Cox regression.

The practitioner who is trying to determine sample size is faced with a dilemma as it usually won't be apparent until the data have been collected that maximum likelihood is a poor choice for that set of data, but the sample size has been determined under the assumption that maximum likelihood will be used and sample size formulas have not been presented in the literature for alternative estimation methods such as those of Firth (1993).

As indicated in Section 4.2.1, the method given by Vaeth and Skovlund (2004) can be used in Cox regression (when Cox regression can be appropriately used, that is). This is not the method that is used in the software discussed in this book and therefore won't be illustrated here.

Sample size determination for Cox regression is available in Stata and PASS. For Stata, this is based on the method of Hsieh and Lavori (2000), which reduces to the method of Schoenfeld (1983) for a binary covariate. This is discussed and

illustrated in Section 9.2. The procedure in PASS assumes use of the model given in Eq. (5.8) and also uses the formula given by Hsieh and Lavori (2000), which is the same type of inflation formula given by Hsieh et al. (1998) for logistic regression. That is, the sample size for testing the null hypothesis that $\beta_1 = 0$ is determined in part by the correlation that X_1 has with the other covariates, with the formula given by

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2}{(1 - R^2)\sigma^2 B^2}$$

with B being the value of β_1 that an experimenter wishes to detect.

Whether or not the inflation factor is appropriate can perhaps be questioned, just as Hosmer and Lemeshow (2000) questioned the similar formula of Hsieh et al. (1998).

■ EXAMPLE 5.3

Assume that we wish to determine the sample size for testing a parameter in a Cox regression model, with $\alpha = .05$ for a two-sided test and the desired power is .90. With the null hypothesis being that the parameter value is zero, we want to detect a value of 0.5 when the standard deviation of the corresponding covariate is 1.5, there are no censored observations, and the value of R^2 when the covariate is regressed against the other covariates in the model (using linear regression) is 0.30. The output from PASS shows that the required sample size is 27 and the power is .903. If the computation had been performed by hand, the result would have been

$$\begin{aligned} n &= \frac{(Z_{\alpha/2} + Z_\beta)^2}{(1 - R^2)\sigma^2 B^2} \\ &= \frac{(1.96 + 1.28155)^2}{(1 - .3)(2.25)(0.5)^2} \\ &= 26.686 \end{aligned}$$

so that $n = 27$ would be used, in agreement with the PASS output. ■

5.4 POISSON REGRESSION

Poisson regression is used when regression is applied to count data and the dependent variable is distributed as approximately a Poisson distribution. As with logistic regression, the estimation procedure is iterative. The parameter estimates are usually obtained as maximum likelihood estimates, using iteratively reweighted least squares. Although exact Poisson regression is an alternative to

maximum likelihood, and will often be preferable for small samples, there is apparently no software that will determine sample size for exact Poisson regression. Some software does have capability for using exact Poisson regression, however, such as Stata.

The model for Poisson regression is

$$\log(\text{Rate}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

so

$$\text{Rate} = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \tag{5.10}$$

and the model for the count variable is then n times the right-hand side, with n as usual denoting the sample size, which could be determined using software unless the user is already set on a sample size. That is,

$$\begin{aligned}\text{Count} = Y &= n\left[\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)\right] \\ &= \exp[\ln(n) + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k]\end{aligned}$$

Signorini (1991) proposed a method for determining sample size and that method is incorporated into G*Power and PASS, with these being the only major sample size determination software with capability for Poisson regression. Specificially, power or sample size is determined for the model in Eq. (5.10) for the null hypothesis $\beta_1 = 0$ versus the alternative hypothesis $\beta_1 = B1$ (using the notation in PASS).

When X_1 is the only regression variable in the model, the minimum sample size, n, is determined from (Signorini, 1991)

$$n \geq \phi \frac{\left[z_{\alpha/2}\sqrt{\text{Var}(\hat{\beta}_1|\beta_1 = 0)} + z_\beta \sqrt{\text{Var}(\hat{\beta}_1|\beta_1 = B1)}\right]^2}{\mu_T \exp(\beta_{10}) B1^2} \tag{5.11}$$

where ϕ denotes a measure of overdispersion, μ_T is the mean exposure time, β_{10} denotes the value of β_1 under the null hypothesis, z is the standard normal deviate, and this is for a two-sided test. For a one-sided test, $z_{\alpha/2}$ would be replaced by z_α.

To illustrate, the null and alternative hypotheses are generally expressed in terms of $\exp(\beta_1)$ rather than β_1 itself. Thus, $\exp(\beta_1) = 1$ under the null hypothesis that $\beta_1 = 0$, and we will assume that we want to detect, with a one-sided test, a reduction to $\exp(B1) = 0.7$, which implies that $B1 = -0.356675$. $\text{Var}(\hat{\beta}_1) = e^{-\beta_1^2/2}$ is the asymptotic variance under the assumption that X has a standard normal distribution, with the variance form for other distributions of X given in Table 1 of Signorini (1991). Thus, for this distributional assumption, $\text{Var}(\hat{\beta}_1|\beta_1 = 0) = e^0 = 1$ and $\text{Var}(\hat{\beta}_1|\beta_1 = -0.356675) = \exp[(-0.5)(-0.356675)^2] = 0.938372$.

Thus, for $\alpha = .05$ and power $= .95$ and assuming no overdispersion (so that $\phi = 1$), the necessary minimum sample size is

$$\begin{aligned} n &= \phi \frac{\left[z_\alpha \sqrt{\text{Var}(\hat{\beta}_1 | \beta_1 = 0)} + z_\beta \sqrt{\text{Var}(\hat{\beta}_1 | \beta_1 = B1)}\right]^2}{\mu_T \exp(\beta_{10}) B1^2} \\ &= \frac{\left[1.645\sqrt{1} + 1.645\sqrt{(0.938372)}\right]^2}{(1)(1)(-0.356675)^2} \\ &= 82.44 \end{aligned} \tag{5.12}$$

Thus, the sample size would be rounded up to $n = 83$, in agreement with the solution given by PASS. Table 2 of Signorini (1991) gives $n = 82$, presumably because there was rounding to the nearest integer, but as has been stated previously, the rounding must be up to ensure that the power is at least equal to the desired power. For those readers who wish to consult Signorini (1991), it is worth noting that the same thing happens when the alternative hypothesis is $\exp(B1) = 0.5$, for example, as use of Eq. (5.12) leads to $n = 20.05$, so that Table 2 of Signorini (1991) gives $n = 20$, whereas PASS properly gives $n = 21$. Such one-unit differences are really inconsequential, however, because an asymptotic variance expression is being used, which would not apply to such small samples. So both solutions are thus potentially incorrect. It should also be noted that Signorini (1991) did not discuss one-sided and two-sided tests, but simply gave Eq. (5.11), which obviously applies to two-sided tests, whereas the entries in Table 2 of Signorini (1991) are obviously for one-sided tests. Thus, there is the potential for confusion.

When there are other regression variables in the model, PASS, for example, follows the lead of Hsieh et al. (1998) for multiple logistic regression and Hsieh and Lavori (2000) for Cox regression with multiple covariates and inflates the sample size for Poisson regression with a single covariate by $1/(1 - R^2)$, with R^2 denoting the square of the correlation between the covariate that corresponds to the parameter that is being tested and the other covariates in the model.

Thus, using the example given earlier in this section, if there were other covariates in the model and $R^2 = .90$, the sample size would be $10(82.44) = 824.4$, so $n = 825$ would be used, which is the solution given by PASS. The latter uses the formula

$$n = \phi \frac{\left[z_{\alpha/2} \sqrt{\text{Var}(\hat{\beta}_1 | \beta_1 = 0)} + z_\beta \sqrt{\text{Var}(\hat{\beta}_1 | \beta_1 = B1)}\right]^2}{\mu_T \exp(\beta_0) B1^2 (1 - R^2)}$$

for a two-sided test, with R^2 denoting the square of the multiple correlation coefficient when the variable of interest is regressed on the other regression

variables in the model. It can be observed that this is the formula due to Signorini (1991) given earlier in this section, divided by $(1 - R^2)$.

Thus, the sample size is larger when there are other regression variables in the model since $(1 - R^2) < 1$. Although there has apparently not been any published research to confirm this, this inflation factor may be suspect for use in Poisson regression with multiple covariates, just as it is for logistic regression, as discussed in Section 5.2.2.

Just as was done for logistic regression, Shieh (2001) offered an improvement over the approach given by Signorini (1991) for Poisson regression. Specifically, Shieh's (2001) method incorporated the limiting value of the maximum likelihood estimates of nuisance parameters under the composite null hypothesis. Since it is based on maximum likelihood estimates, the method may not work very well when maximum likelihood is not the best method of estimation, but as indicated previously, this isn't going to be known until the data have been collected.

5.5 NONLINEAR REGRESSION

Determining sample size for nonlinear regression would be a very formidable task because there are so many possible nonlinear regression models. Therefore, it is not surprising that sample size capability for nonlinear regression is not available in any of the leading software (e.g., it is not available in PASS), nor has it apparently been discussed in the literature.

5.6 OTHER TYPES OF REGRESSION MODELS

There are other types of regression models (probit regression, random coefficient regression, etc.) but there is no widely available sample size determination software for such models, nor has there apparently been any research on sample size determination for such models.

5.7 CORRELATION

Correlation is related to linear regression in the following way. If X is a *random* variable, we may properly speak of the *correlation* between X and Y. [Note that the term "correlation" is frequently used (improperly) when the values of X to be used in a regression model are preselected rather than occurring as realizations of a random variable, but the practice is longstanding and not likely to change.]

Correlation refers to the extent that the two random variables are related, with the strength of the relationship measured by the (sample) correlation coefficient, r_{xy}. There are many types of correlation coefficients, whose use depends on whether the variables are continuous, binary, ordinal, and so on. Some of these

are mentioned briefly in Chapter 11, with references that discuss sample size determination.

The most commonly used correlation coefficient, which is for measuring the strength of the linear relationship between two continuous random variables that have a joint bivariate distribution (a seldom checked requirement), is the Pearson sample correlation coefficient. It is computed as

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{5.13}$$

with $S_{xx} = \sum(X - \bar{X})^2$, $S_{xy} = \sum(X - \bar{X})(Y - \bar{Y})$, and $S_{yy} = \sum(Y - \bar{Y})^2$. Formally, the population correlation coefficient, ρ_{xy}, is defined as the covariance between X and Y divided by the standard deviation of X times the standard deviation of Y. When those parameters are estimated by the sample statistics, the result reduces to Eq. (5.13).

The possible values of r_{xy} range from -1 to $+1$. The former represents perfect negative correlation, as would happen if all the points fell on a line with a negative slope, and the latter represents perfect positive correlation. In regression there is nothing really "negative" about negative correlation, though, as a strong negative correlation is just as valuable as a strong positive correlation as far as estimation and prediction are concerned, although usually the Xs in a regression model are positively correlated with Y. A zero correlation between X and Y would signify that it would not be meaningful to construct a linear regression equation using that regressor, and the same could be said if the correlation is small.

Although a user could solve for the sample size for a hypothesis test of H_0: $\rho = 0$ against a logical alternative, we would certainly expect there to be a nonzero correlation between Y and any particular X that is being considered for a simple linear regression model. Thus, this would be a somewhat meaningless exercise. (Recall the statement made in Section 1.1 that null hypotheses are almost always false; this would be an example of that.)

What would not be meaningless, however, would be to proceed along the lines of what was discussed in Section 5.1.1 regarding Soper's applet.

Since R^2 discussed in that section equals r_{xy}^2, the discussion there directly applies to correlation. That is, one could decide upon a minimally acceptable value of r_{xy} to allow the user to claim a meaningful (linear) correlation between X and Y, then convert that to R^2 and then to the corresponding effect size and use Soper's applet. To illustrate, assume that 0.7 is chosen as the dividing line between an acceptable and unacceptable correlation. Thus, $R^2 = .49$ and the effect size, f^2, is 0.96. With Soper's applet we obtain $n = 11$ if the power is specified as .80 and $\alpha = .05$; we obtain $n = 14$ if the power is .90.

When the user-written command `sampsi_rho` for Stata is used, a sample size of $n = 11.22$ results when 0.7 is specified as the value of the correlation coefficient under the alternative hypothesis and power of .80 and a one-sided test

are specified. When the power is increased to .90, Stata gives 14.38 as the sample size. (The actual powers are .832 and .914, respectively.) These are also, oddly, the values of n, to the nearest integer, that result when Power and Precision is used and a *two*-sided test is specified. When a one-sided test is specified, the sample sizes that the software gives are 9 and 11, respectively, for powers of .80 of .90, respectively. PASS, which uses theoretical results given by Guenther (1977), gives 13 and 17, respectively, for these two power values when a two-sided test is used and 11 and 14 for a one-sided test, with nQuery giving 14 and 17 for the two-sided test and 11 and 14 for the one-sided test. The corresponding results given by G*Power are 14 and 17 and 11 and 13, respectively, thus not agreeing with either nQuery or PASS. The disagreement is small, although the percentage disagreement is not small. Lenth's applet agrees with PASS without giving the user the option of choosing between a one-sided and two-sided test. It is not clear why the results given by Power and Precision for a two-sided test are off more than slightly.

Among other applets, the one at `http://www.quantitativeskills.com/sisa/statistics/corrhlp.htm` can be used for both two-sided tests and one-sided tests and, without specific additional input, gives necessary sample sizes for all combinations of $\alpha = .10, .05, .01$, and .001, and power values of .6, .7, .8, and .9. The results agree with the aforementioned results given by PASS.

When there is more than one independent variable that is of interest relative to a dependent variable, determining sample size to detect a certain minimum (nonzero) value of a multiple correlation coefficient may be of interest. Gatsonis and Sampson (1989) addressed this issue and presented tables for sample size determination.

5.7.1 Confidence Intervals

Of course, a confidence interval for ρ of a desired width could also be constructed, and would be a more practical approach because testing a null hypothesis of zero doesn't accomplish very much. To illustrate, assume as in the preceding example that $\widehat{\rho} = .7$ and we want a 95% two-sided confidence interval to be of width 0.1. In order to accomplish this, the sample size must be 404, according to PASS. Since that is a very large sample, a width of .20 might have to be used. This requires a sample size of 105, which might still be larger than practical in certain applications.

Interestingly, nQuery provides a hypothesis test for ρ that can be either one-sided or two-sided, but sample size can be determined for only a one-sided confidence bound. Because of asymmetry of the distribution of r_{xy}, the required sample size will depend on whether an upper limit or lower limit is desired. For example, if $\widehat{\rho} = .7$ and an upper limit of .8 is specified, nQuery gives $n = 54$, but $n = 93$ results if a lower limit of .6 is specified. These results are in agreement with the results given by PASS, which has the capability for one-sided confidence bounds in addition to two-sided intervals.

Although G*Power has the capability for testing several different types of correlation coefficients, no confidence intervals of any type are provided. Similarly, although MINITAB has the capability for sample size determination for some types of confidence intervals, this does not include a confidence interval for ρ. Power and Precision does not have the capability to compute sample sizes for confidence interval widths directly, but a user can enter a number for the population correlation and the limits are displayed. Thus, the user could use some directed trial and error to arrive at a sample size for a confidence interval of a desired width as a by-product of the hypothesis tests. Unfortunately, when a one-sided hypothesis test is specified, the output gives the limits for a two-sided confidence interval, so the hypothesis test and confidence interval results are not connected.

5.7.2 Intraclass Correlation

There are different types of intraclass correlation coefficients, with intraclass correlation originally for paired measurements, such as measurements made on the same subject. This is the way that intraclass correlation is incorporated into PASS, with sample size determined relative to the hypothesis test that the intraclass correlation is equal to a specified value.

For example, if the null hypothesis of no intraclass correlation is tested against the alternative hypothesis that the intraclass correlation is 0.6, the number of subjects must be input into PASS, which will give the number of observations to use per subject. (In a random effects linear model, with treatments the random factor, this is the number of treatments rather than the number of people.) When $\alpha = .05$, power $= .90$, and 5 is entered for the number of subjects, PASS gives $n = 7$ observations per subject. This is a one-sided test, as a two-sided test is not an option, unlike the options that are available for testing a Pearson correlation coefficient, which include a two-sided test and an upper and lower one-sided test. See Bonett (2002) for sample size determination for estimation of intraclass correlation coefficients.

5.7.3 Two Correlations

It is sometimes of interest to test whether two correlation coefficients are equal, such as determining if two different methods of school instruction correlate equally with student test performance. This is completely analogous to testing whether the slopes of two simple linear regression models are equal because if both X and Y are put in correlation form, the (single) regression coefficient in the equation is the linear (Pearson) correlation between X and Y. So testing for the equality of two correlation coefficients would indirectly test the equality of the slopes of the two simple linear regression models.

Segmented regression refers to the fitting of two or more linear regression models to a set of data, with the different models fit for nonoverlapping ranges

of the data. For example, a simple linear regression model might be fit for $12 \leq X \leq 18$ and a different simple linear regression model fit for $19 \leq X \leq 30$. This decision might be based on subject-matter knowledge. If the assumption of the need for two models is not well founded, however, a test of the equality of the two correlation coefficients might be performed, with a decision made to fit two models only if the sample correlations differ by at least a specified amount. The task is then to determine the sample size for each group, after specifying the desired power and significance level.

Power and Precision and PASS are two software packages that can be used for solving for sample size. Of course, the two correlation coefficients should be reasonably close to one; otherwise, neither simple linear regression model would provide a good fit to the data.

So let's specify one correlation at .80 and the other at .90, with this difference of .10 being the minimum difference that we want to detect. Using either Power and Precision or PASS, the sample size for each group is found to be 154, so the total number of observations will be 308. Such a large number may not be obtainable, however. If so and if power of .80 is acceptable, 116 observations would be required for each group, for a total of 232. Of course, larger sample sizes would be required if the difference between the two correlations was less than .10. Depending on the type of application, the required sample sizes might not be attainable, even for a power of .80 and a difference of .10.

The applet mentioned in Section 5.7 gives results that agree with the results given by PASS and Power and Precision for this example and provides sample sizes for one- and two-sided tests of the null and alternative hypotheses for all combinations of $\alpha = .10, .05, .01, .001$ and power $= .6, .7, .8$, and .9.

5.8 SOFTWARE

Overall, software for sample size determination for regression models is somewhat of a problem. First, software is not available for all types of regression models and there are some problems with existing software, both in terms of different software producing different solutions, as noted in Section 5.2.1.1 for logistic regression because different algorithms are being used, and in terms of what the user must input. It is best not to focus on individual regression coefficients in multiple linear regression and multiple logistic regression, but rather determine sample size by using a measure of model worth, such as R^2. Although this is straightforward in linear regression because there is one commonly used form of R^2, there are multiple forms of R^2 that have been proposed for logistic regression.

Another problem with software is the absence of capability for determining sample size when a necessary alternative to maximum likelihood should be used, such as the method proposed by Firth (1993).

5.9 SUMMARY

Sample size determination for regression models is difficult (and almost inadvisable when the focus is on individual regression coefficients) because of the problems involved in trying to interpret regression coefficients. A new method that involves the Wetz criterion was presented for linear regression but there is no way to avoid guesswork unless there is prior information available. Another problem is that the common use of maximum likelihood as the method of estimation can create problems in logistic regression (binary and polytomous), as well as in Cox regression. These problems can overshadow problems caused by using a sample size that is too small or too large when these methods are employed. Furthermore, even in simple linear regression there are quantities that must be specified which will generally be unknown and may be difficult to estimate, such as the variance of the predictor when it is a random variable, as discussed in Section 5.1.1. Problems were noted with a proposed method of determining sample size for multiple logistic regression, and although a simple rule-of-thumb might be preferable, analogous to the 10:1 rule-of-thumb for multiple linear regression, it would be difficult to construct a simple rule motivated by the results of Peduzzi et al. (1996) because it may be difficult to estimate how many events there should be for combinations of planned or anticipated values of the predictor variables. Of course, there is also parameter estimation error in virtually all practical applications of sample size determination and Taylor and Muller (1996) have looked at the bias this creates for linear models.

REFERENCES

Adcock, C. J. (1997). Sample size determination: A review. *The Statistician*, **46**(2), 261–283.

Alam, M. K., M. B. Rao, and F.-C. Cheng (2010). Sample size determination in logistic regression. *Sankya*, **72B**, Part 1, 58–75.

Algina, J. and S. Olejnik (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research*, **35**(1), 119–136.

Ammann, R.A. (2004). Correspondence. Bone Marrow Transplantation, **34**, 277–278. (Available at `http://www.cytel.com/Papers/2004.08_Letter_nature.pdf`.)

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine*, **21**(9), 1331–1335.

Brownlee, K. A. (1960). *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley.

Bull, S. B. (1993). Sample size and power determination for a binary outcome and an ordinal response when logistic regression is planned. *American Journal of Epidemiology*, **137**, 676–684.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. New York: Routledge Academic.

Cook, R. D. and S. Weisberg (1999). *Applied Regression Including Computing and Graphics.* New York: Wiley.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Chapman and Hall.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Dattalo, P. (2008). *Determining Sample Size: Balancing Power, Precision, and Practicality*. New York: Oxford University Press.

Demidenko, E. (2007). Sample size determination for logistic regression revisited. *Statistics in Medicine*, **26**(18), 3385–3397.

Demidenko, E. (2008). Sample size and optimal design for logistic regression with binary infection. *Statistics in Medicine*, **27**, 36–46.

Draper, N. R. and H. Smith (1981). *Applied Regression Analysis*, 2nd edition. New York: Wiley. (The current edition is the 3rd edition, 1998.)

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

Flack, V. F. and T. L. Eudey (1993). Sample size determinations using logistic regression with pilot data. *Statistics in Medicine*, **12**, 1079–1084.

Gatsonis, C. and A. R. Sampson (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin*, **106**(3), 516–524.

Gorman, J. W. and R. J. Toman (1966). Selection of variables for fitting equations to data. *Technometrics*, **8**, 27–51.

Guenther, W. C. (1977). Desk calculation of probabilities for the distribution of the sample correlation coefficient. *The American Statistician*, **31**(1), 45–48.

Halinksi, R. S. and L. S. Feldt (1970). The selection of variables in multiple regression analyses. *Journal of Educational Measurement*, **7**(3), 151–158.

Hosmer, D. W. Jr. and S. Lemeshow (2000). *Applied Logistic Regression*, 2nd edition. New York: Wiley.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine*, **8**, 795–802.

Hsieh, F. Y. and P. W. Lavori (2000). Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled Clinical Trials*, **21**, 552–560.

Hsieh, F. Y., D.A. Bloch, and M. D. Larsen (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, **17**, 1623–1634.

Kelley, K. and S. E. Maxwell (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, **8**, 305–321. (Available at `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.5835&rep=rep1&type.pdf`.)

Krishnamoorthy, K. and Y. Xia (2008). Sample size calculation for estimating or testing a nonzero squared multiple correlation coefficient. *Multivariate Behavioral Research*, **43**(3), 382–410.

King, E. N. and Ryan, T. P. (2002). A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician*, **56**, 163–170.

Kodell, R. L., S. Y. Lensing, R. D. Landes, K. S. Kumar, and M. Hauer-Jensen (2010). Determination of sample sizes for demonstrating efficacy of radiation countermeasures. *Biometrics*, **66**, 239–248.

Kraemer, H. C. and S. Thiemann (1987). *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage Publications.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, **55**, 187–193.

Lindsey, J. K. (1997). Exact sample size calculations for exponential family models. *The Statistician*, **46**(2), 231–237.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

Maxwell, S. E. (2000). Sample size and multiple regression analsyis. *Psychological Methods*, **5**, 434–458.

Mehta, C. R. and N. R. Patel (1995). Exact logistic regression: Theory and examples. *Statistics in Medicine*, **14**, 2143–2160.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, **54**, 17–24.

Mendoza, J. L. and K. L. Stafford (2001). Confidence intervals, power calculation, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables. *Educational and Psychological Measurement*, **61**(4), 650–667.

Miller, D. E. and J. T. Kunce (1973). Prediction and statistical overkill revisited. *Measurement and Evaluation in Guidance*, **6**(3), 157–163.

Müller, K. E. and V. B. Pasour (1997). Bias in linear model power and sample size due to estimating variance. *Communications in Statistics—Theory and Methods*, **26**(4), 839–852.

Newson, R. (2004). Generalized power calculations for generalized linear models and more. *The Stata Journal*, **4**(4), 379–401.

Novikov, I., N. Fund, and L. S. Freedman (2010). A modified approach to estimating sample size for simple logistic regression with one continuous covariate. *Statistics in Medicine*, **29**(1), 97–107.

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, **49**(12), 1373–1379.

Ryan, T. P. (2009a). *Modern Regression Methods*, 2nd edition. Hoboken, NJ: Wiley.

Ryan, T. P. (2009b). Maximum Likelihood. Course developed for The Institute for Statistics Education.

Ryan, T. P. and W. H. Woodall (2005). The most-cited statistical papers. *Journal of Applied Statistics*, **32**(5), 1–14.

Sawyer, R. (1983). Determining minimum sample sizes for multiple regression grade prediction equations for colleges. *American Statistical Association Proceedings of the Social Statistics Section*, pp. 379–384. (Also as Research Report No. 83, American College Testing Program, February 1984.)

Schoenfeld, D. A. (1983). Sample size formula for the proportional-hazards regression model. *Biometrics*, **39**, 499–503.

Schouten, H. J. A. (1999). Sample size formula with a continuous outcome for unequal group sizes and unequal variances. *Statistics in Medicine*, **18**, 87–91.

Self, S. G. and R. H. Mauritsen (1988). Power/sample size calculations for generalized linear models. *Biometrics*, **44**, 79–86.

Self, S. G., R. H. Mauritsen, and J. Ohara (1992). Power calculations for likelihood ratio tests. *Biometrics*, **48**, 31–39.

Sharma, D. R. (2006). Logistic regression, measures of explained variation, and the base rate problem. Ph.D. dissertation. Department of Statistics, Florida State University. (Available at http://etd.lib.fsu.edu/theses/available/etd-06292006-153249/unrestricted/dissertation_drs.pdf.)

Shieh, G. (2000). On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics*, **56**, 1192–1196.

Shieh, G. (2001). Sample size calculations for logistic and Poisson regression models. *Biometrika*, **88**(4), 1193–1199.

Signorini, D. F. (1991). Sample size for Poisson regression. *Biometrika*, **78**(2), 446–450.

Simon, S. (2008). Sample size for an ordinal outcome. (Electronic resource: www.pmean.com/04/ordinalLogistic.html.)

Taylor, D. J. and K. E. Müller (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics—Theory and Methods*, **25**(7), 1595–1610.

Taylor, A. B., S. G. West, and L. S. Aiken (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and Psychological Measurement*, **66**(2), 228–239.

Thigpen, C. C. (1987). A sample size problem in simple linear regression. *The American Statistician*, **41**, 214–215.

Tosteson, T. D., J. S. Buzas, E. Demidenko, and M. Karagas (2003). Power and sample size calculations for generalized regression models with covariate measurement error. *Statistics in Medicine*, **22**(7), 1069–1082.

Vaeth, M. and E. Skovlund (2004). A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Statistics in Medicine*, **23**(11), 1781–1792.

Vaughn, C. and S. Guzy (2002). Redesigning experiments with polychotomous logistic regression: A power computation application. Paper 26-27. SUGI27.

Walters, S. J. (2004). Sample size and power estimation for studies with health related quality of life outcomes: A comparison of four methods using the SF-36. *Health and Quality of Life Outcomes*, **2**, 26. (Open access; available at www.hqlo.com/content/2/1/26 and www.hqlo.com/content/pdf/1477-7525-2-26.pdf.)

Wetz, J. M. (1964). Criteria for judging adequacy of estimation by an approximating response function. Ph.D. thesis. Department of Statistics, University of Wisconsin.

Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in Medicine*, **12**, 2257–2271. (Erratum: April 30, 1994; **13**(8), 871.)

Whittemore, A. S. (1981). Sample size for logistic regression with small response probability. *Journal of the American Statistical Association*, **76**, 27–32.

EXERCISES

5.1. Verify Eq. (5.2).

5.2. Use the Demidenko applet at http://www.dartmouth.edu/~eugened/power-samplesize.php to determine the required sample size for a one-variable logistic regression problem with a binary covariate, such that 40% of the covariate values are equal to 1 and when the covariate value is 0,

60% of the response values are equal to 1, and the detectable odds ratio is 2.5. Use a significance level of .05 and power of .90.

5.3. A scientist comes to you and asks you to determine the sample size that she should obtain when linear regression is to be used. There is to be just a single predictor so this will be simple linear regression. The predictor, however, is a random variable. The scientist tells you that she wants to have a good regression model with $R^2 \geq .75$. She realizes that a very small sample size is not a good idea but there is a sampling cost involved, which is not small. She has heard of the Wetz criterion and believes the criterion should be met. Assume the use of $\alpha = .05$.

5.4. An experimenter makes the following statement: "Because of the difficulty in determining sample size in simple linear regression, especially since my predictor will be random and samples in my field are costly, I am simply going to use the Draper and Smith rule-of-thumb and use a sample size of 10. Then I will enter the sample quantities from my experiment into PASS, including the sample size of 10 that I used, and see what power my experiment had, as given by PASS." Critique that statement.

5.5. How would you advise a researcher to determine sample size for multiple binary logistic regression?

5.6. The regression methods discussed in the chapter all rely on maximum likelihood as the method of estimation (ordinary least squares is equivalent to maximum likelihood under the assumption of normality), even though maximum likelihood could be a poor choice, depending on the data (King and Ryan, 2002). Of course, when the sample size is being determined, there are no data, unless there are data available from a pilot study or similar study. Therefore, what is the first step that you would recommend for a user once the data have been collected?