

NOTES ON LOGISTIC REGRESSION

1. The Logit Model.

These notes introduce a special regression model for analyzing *dichotomous* dependent variables, such as voter opinions (“Yes” vs “No”) or test results (“Pass” vs “Fail”), typically represented by a binary random variable $Y = 0, 1$. To see the need for a special model, note that for any explanatory variables, (x_1, \dots, x_k) , the *standard linear model*

$$(1) \quad Y = \beta_0 + \sum_{j=1}^k \beta_j x_j + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

assumes that Y is normally distributed (and hence *cannot* be binary). A much better way to model this data is to assume that the variables (x_1, \dots, x_k) influence the probability that $Y = 1$ through some appropriate transform of the linear function in (1). The simplest function of this type is the *logit function*, which assumes in particular that:

$$(2) \quad \Pr(Y = 1 | x_1, \dots, x_k) = \frac{\exp\left(\beta_0 + \sum_{j=1}^k \beta_j x_j\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^k \beta_j x_j\right)}$$

and hence that

$$(3) \quad \Pr(Y = 0 | x_1, \dots, x_k) = 1 - \frac{\exp\left(\beta_0 + \sum_{j=1}^k \beta_j x_j\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^k \beta_j x_j\right)} = \frac{1}{1 + \exp\left(\beta_0 + \sum_{j=1}^k \beta_j x_j\right)}$$

This is called the (*linear*) *logit model* of Y with respect to (x_1, \dots, x_k) .

2. Maximum-Likelihood Estimation of Parameters

If for any given set of data $(y_i, x_{1i}, \dots, x_{ki}), i = 1, \dots, n$ it is assumed that the n samples are independent, then the joint probability of the observed values (y_1, \dots, y_n) is simply

$$(4) \quad \Pr(y_1, \dots, y_n) = \prod_{i=1}^n \Pr(Y_i = y_i | x_{1i}, \dots, x_{ki})$$

By substituting the appropriate expressions [either (2) or (3)] into each factor of the right hand side of (4), one can write this probability as an explicit function of the unknown parameters $(\beta_0, \beta_1, \dots, \beta_k)$. The resulting function is called the *likelihood function* for $(\beta_0, \beta_1, \dots, \beta_k)$ given (y_1, \dots, y_n) , and can be written as

$$L(\beta_0, \beta_1, \dots, \beta_k | y_1, \dots, y_n) = \prod_{i=1}^n \left\{ \left(\frac{\exp(\beta_0 + \sum_{j=1}^k \beta_j x_j)}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_j)} \right)^{y_i} \left(\frac{1}{1 + \exp(\beta_0 + \sum_{j=1}^k \beta_j x_j)} \right)^{1-y_i} \right\}$$

Note that whenever $y_i = 1$, it follows that $1 - y_i = 0$ and hence that the term in brackets becomes (2). Conversely, when $y_i = 0$ it follows that $1 - y_i = 1$ and hence that this term becomes (3).

Next observe that by maximizing L with respect to the unknown parameters $(\beta_0, \beta_1, \dots, \beta_k)$, one must obtain the parameter values $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ under which the observed outcomes (y_1, \dots, y_n) are most likely to occur. These parameter values $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ are called the *maximum-likelihood estimates* of $(\beta_0, \beta_1, \dots, \beta_k)$ given (y_1, \dots, y_n) . In the present case this maximum-likelihood estimation procedure is designated as *logistic regression*, and is available in JMPIN.

3. An Application using JMPIN

To illustrate logistic regression in JMPIN, we employ the weather data set in the JMPIN file, **Rain_Logit.jmp**. This contains weather data collected at the same location for each 30 days. Here the dichotomous variable Y is designated as **Rained**, and has nominal values “Rainy” vs “Dry”. The nominal categorization of this variable is denoted in JMPIN by the red icon next to **Rained** in list of columns on the left side of the data window. The two explanatory variables (x_1, x_2) correspond to barometric pressure (**Pressure**) and relative humidity (**Humidity**) respectively.

To run a logistic regression of **Rained** on (**Pressure, Humidity**), simply proceed as with any multiple regression by using the **Fit Model** option, where **Y** is specified as **Rained**, and where **Pressure** and **Humidity** are added as the explanatory variables. Notice that the **Personality** box in the upper right corner of the **Fit Model** window now reads “Nominal Logistic” rather than “Standard Least Squares”. JMPIN has automatically detected the presence of a nominal variable as **Y** and selected logistic regression as the default option for this case.

When you click **Run Model**, you will now see that the output results of this maximum-likelihood estimation procedure are substantially different from those of multiple

regression. First notice that there are *no* graphical outputs. (If you run a logistic regression with one explanatory variable using **Fit Y by X** you will see a graphical output. However, this graph is difficult to interpret, and is best ignored.)

3.1 Parameter Estimation

Start by looking at the **Parameter Estimates** window, shown below:

Parameter Estimates				
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-612.89105	299.1737	4.20	0.0405
Pressur	21.0482199	10.243164	4.22	0.0399
Humidity	-0.0899832	0.0480532	3.51	0.0611
For log odds of Dry/Rainy				

Here the **Estimate** ($\hat{\beta}_j$) and **Std Error** (s_j) look exactly like those for multiple regression. However, these $\hat{\beta}_j$'s are *maximum-likelihood* estimates and not *least-squares* estimates. Moreover, observe that the *signs* of these beta coefficients appear to be *wrong*: **Rain** should be more likely when **Pressure** *decreases* and/or **Humidity** *increases*. The difficulty here is that JMPIN has set 1 = "Dry" and 0 = "Rainy". In making its choice, JMPIN automatically chooses the *first* nominal value (alphabetically) of the dependent variable as the *event value* ($Y = 1$) in computing the probability formulas in (2) and (3) above. In the present case, "Dry" is the event value so that (2) is computed for "Dry", as can be seen by the fact that the linear formula reported in the columns above is **Lin[Dry]** rather than **Lin[Rainy]**. If you wish to ensure that (2) is computed for the event of most interest, you must label this event so that it is first alphabetically. For example, in the file Rain_Logit.jmp, a new variable, **Rained_New**, has been constructed with nominal values, **a_Rainy** and **b_Dry**. If the logistic regression is run on this variable, then JMPIN will make **a_Rainy** the event value. Try it.

3.1.1. Interpretation of Beta Values. Unlike linear regression, one cannot interpret the beta values, β_j , as simple partial derivatives of $E(Y | x_1, \dots, x_k)$ with respect to the x_j 's. However, one can interpret them in terms of "odds ratios". First, observe that if the *odds ratio* of $Y = 1$ vs $Y = 0$ for a given set of attributes (x_1, \dots, x_k) is denoted by

$$(5) \quad R(x_1, \dots, x_k) = \frac{\Pr(Y = 1 | x_1, \dots, x_k)}{\Pr(Y = 0 | x_1, \dots, x_k)}$$

[so that for example, $\Pr(Y = 1|x_1, \dots, x_k) = .75$ implies $R(x_1, \dots, x_k) = .75 / .25 = 3$, and hence that odds of the relevant event $Y = 1$ occurring are “3 to 1”], then by (2) and (3) we see that

$$(6) \quad R(x_1, \dots, x_k) = \exp\left(\beta_0 + \sum_{j=1}^k \beta_j x_j\right)$$

Hence if any explanatory variable, x_j , is incremented by one unit, then we see that

$$(7) \quad \begin{aligned} \frac{R(x_1, \dots, x_i + 1, \dots, x_k)}{R(x_1, \dots, x_i, \dots, x_k)} &= \frac{\exp\left(\beta_0 + \beta_j(x_j + 1) + \sum_{l \neq j} \beta_l x_l\right)}{\exp\left(\beta_0 + \beta_j x_j + \sum_{l \neq j} \beta_l x_l\right)} \\ &= \exp\left[\left(\beta_0 + \beta_j(x_j + 1) + \sum_{l \neq j} \beta_l x_l\right) - \left(\beta_0 + \beta_j x_j + \sum_{l \neq j} \beta_l x_l\right)\right] \\ &= \exp(\beta_j) \end{aligned}$$

Thus if $\exp(\beta_j) = 2$ then this would mean that incrementing x_j by one unit would double the odds of the event $Y = 1$ occurring. It is this interpretation of β_j that is typically adopted. However, the value of “one unit” depends very much on the units of measure. For example, in the present example, a full inch of mercury in barometric pressure is an enormous pressure change. So if we use the dependent variable **Rained_New**, and take **a_Rainy** to be the relevant event, then the value $\beta_j = -21.05$ yields such a drastic reduction in the odds ratio [$\exp(\beta_j) = \exp(-21.05) \approx 10^{-9}$] that it is difficult to interpret. However, if one *redefines* the appropriate increment in x_j to be of any size α (rather than 1), then the right hand side of (7) becomes $\exp(\alpha \cdot \beta_j)$. So for example, if it is more meaningful to consider increments in barometric pressure of say a tenth of an inch of mercury, then (7) would yield the value $\exp(0.1 \cdot \beta_j) = \exp(-2.105) \approx 1/8$. For *inverse* relationships of this type, a more intuitive interpretation might then be to say the model predicts that a *fall* in the barometric pressure of a tenth of an inch will lead to an *eight-fold increase* in the odds of rain.

Note finally that the convention in JMPIN is to use the *range* of x_j in the given data set as the relevant “increment” size α . In particular, if one right clicks on the top bar (**Nominal Logistic Fit**) and selects **Odds Ratio**, then the value observed for **Pressure** is 0.00000174. To understand this value, use the calculator to find the maximum and minimum pressure values (**Formula** → **Statistical** → **Col Minimum** and **Col Maximum**). Here one obtains the values $\text{Pressure}_{\max} = 29.71$ and $\text{Pressure}_{\min} = 29.08$. Hence the JMPIN value is given by $\exp\left[(29.71 - 29.08)(-21.05)\right] = 0.00000174$.

3.1.2. Standard Errors and P-Values for Beta Estimates. Next we consider the standard errors, s_j , which in this case are now estimates of the *asymptotic* standard deviations of the parameter estimates (as the sample size approaches infinity). The **ChiSquare** value for each is simply the square of the standardized value $(\hat{\beta}_j/s_j)$ [under the null hypothesis that $\hat{\beta}_j = 0$]. As an illustration, observe that for the **Pressure** variable, **ChiSquare** = 4.22 = $(21.05/10.24)^2 = (\hat{\beta}_1/s_1)^2$. The associated Chi-Square random variable is thus simply the square of a normal variable, with distribution given by Table A.7 in Devore (p.727). In JMPIN the appropriate P-Value can be calculated more explicitly from the Chi-Square distribution (with one degree of freedom) as follows:

$$(8) \quad 1 - \text{ChiSquare Distribution}(4.22, 1) = 0.0399$$

Thus the value is reported above as simply **Prob>ChiSq = .0399**. [Note that is the same as the *limiting* P-Values for multiple regression, **Prob>|z|** (i.e., with **z** replacing **t** in the limit). In the present case, $z = \hat{\beta}_1/s_1 = 21.05/10.24$ yields the normal P-Value, $2 \cdot \Phi(-z) = .0399$.] In the present example, these results suggest that both **Pressure** and **Humidity** are reasonably good indicators of whether or not it **Rained**, with **Pressure** clearly being the more significant of the two. [It should be noted however that there is a serious question as to whether these *asymptotic P-Values* are appropriate for small samples. Since the usual *t*-approximation to the distribution of $\hat{\beta}_j/s_j$ is *not valid* for these maximum-likelihood estimates, the convention in most standard software (including JMPIN) is simply to use the asymptotic normal P-Values for all sample sizes. A discussion of alternative “Quasi t-Tests” can be found in Ben-Akiva, M. and S. Lerman, *Discrete Choice Analysis*, MIT Press, 1985, p.26.]

3.2 Goodness-of-Fit Measures

Next look at the **Whole Model Test** window, shown below:

Whole Model Test				
Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	10.339797	2	20.67959	<.0001
Full	7.986132			
Reduced	18.325929			
RSquare (U)		0.5642		
Observations (or Sum Wgts)		30		
Converged by Gradient				

The basic “Whole Model Test” is conceptually similar to that for multiple regression, except that rather than look at the distribution of “explained” over “unexplained” variation, $R^2/(1-R^2)$, the test is based on a comparison of the “Full” maximum-likelihood value, $L_1 = L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k | y_1, \dots, y_n)$, achieved by the logit model with the “Reduced” maximum-likelihood value, L_0 , obtained under the *null hypothesis*, H_0 , that $\beta_j = 0$ for each variable $j = 1, \dots, k$. Under H_0 the logit model in (2) is reduced to a simple *Bernoulli* model with probability, $P = \exp(\beta_0)/[1 + \exp(\beta_0)]$. Hence (y_1, \dots, y_n) is taken to be a random sample from this Bernoulli distribution. Since the likelihood of P given (y_1, \dots, y_n) is now of the form:

$$(9) \quad L(P | y_1, \dots, y_n) = \prod_{i=1}^n \{P^{y_i} (1-P)^{1-y_i}\} = P^{\sum y_i} (1-P)^{n-\sum y_i}$$

it follows easily (by solving the first-order condition) that the maximum of (9) is achieved when P equals the sample mean of the y_i 's. Hence the maximum-likelihood estimate of P is simply the *relative-frequency value* :

$$(10) \quad \hat{P} = \frac{1}{n} \sum_{i=1}^n y_i$$

and the corresponding *maximum-likelihood value* achieved under H_0 is given by:

$$(11) \quad L_0 = L(\hat{P} | y_1, \dots, y_n) = \hat{P}^{\sum y_i} (1-\hat{P})^{n-\sum y_i}$$

With these definitions, it can be shown that the test statistic

$$(12) \quad \chi^2 = 2[\ln(L_1) - \ln(L_0)]$$

is asymptotically Chi-Square distributed under H_0 with *two* degrees of freedom – one degree of freedom for each parameter set equal to zero in H_0 (in this case β_1 and β_2). The **Whole Model Test** above is then based on this statistic. In the present case, the JMPIN results show that $\ln(L_1) = -7.986$ and $\ln(L_0) = -18.326$, so that (y_1, \dots, y_n) is far less likely under H_0 than under the full logit model. This is strong evidence against H_0 , with Chi-Square value, $\chi^2 = 2[18.326 - 7.986] = 20.6795$, and corresponding P-Value (calculated in JMPIN):

$$(13) \quad 1 - \text{ChiSquare Distribution}(20.6795, 2) = 0.00003232$$

Thus the value is reported above as simply **Prob>ChiSq < .0001**.

Finally it is of interest to ask whether there is some simple measure of “goodness of fit” in this case which is comparable to R^2 for multiple regression. The short answer is *no*. However, a number of such measures have been proposed. The one reported in JMPIN is the so called **RSquare (U)** value [designated more accurately as the *Likelihood Ratio Index (LRI)* in Green, W.H., *Econometric Analysis*, 2nd ed., 1993, p.651]. This value is given (in the present case) by:

$$(14) \quad LRI = 1 - \frac{\ln(L_1)}{\ln(L_0)} = 1 - \frac{7.986}{18.326} = .5642$$

Notice that since $\ln(L_0) \leq \ln(L_1) < 0$ it follows that $0 < \ln(L_1)/\ln(L_0) \leq 1$, and hence that *LRI* has the same value range as R^2 . Moreover, it is also clear that higher values of *LRI* indicate that the logit model is fitting the data much better than the simple Bernoulli model. But beyond this, there is little that can be said. In particular there is no meaningful notion of “fraction of total variation explained”.

An alternative approach, which is much easier to interpret, is simply to look at how well the model actually predicts the dichotomous outcomes. If the *estimated value* of the probability $\Pr(Y_i = 1 | x_{i1}, \dots, x_{ik})$ for each sample $i = 1, \dots, n$ is denoted by

$$(15) \quad \hat{P}(x_{i1}, \dots, x_{ik}) = \frac{\exp(\hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{ij})}{1 + \exp(\hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{ij})}$$

then it is natural to consider the estimated model as “predicting” $Y_i = 1$ whenever

$$(16) \quad \hat{P}(x_{i1}, \dots, x_{ik}) > \frac{1}{2}$$

If one then defines the *outcome prediction* of the model for each sample $i = 1, \dots, n$ by

$$(17) \quad \hat{y}_i = \begin{cases} 1, & \hat{P}(x_{i1}, \dots, x_{ik}) > \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

then the *prediction success* of the model for sample i can be denoted by

$$(18) \quad S_i = \begin{cases} 1, & y_i = \hat{y}_i \\ 0, & y_i \neq \hat{y}_i \end{cases}$$

With these conventions, a natural measure of goodness of fit for a logistic regression model is given by the *prediction success rate*

$$(19) \quad S = \frac{1}{n} \sum_{i=1}^n S_i$$

This can easily be calculated in JMPIN as follows:

- (i) First, right click on the top bar in the **Logistic Regression** window, and select **Save Probability Formula**.
- (ii) This will add four new columns to the data set (**Lin[a_Rainy]**, **Prob[b_Dry]**, **Prob[a_Rainy]**, **MostLikely Rained_New**). In the present case the two most useful columns are **Prob[a_Rainy]** and **MostLikely Rained_New**. The first gives the estimated probability formula in (15) above, and can be used to obtain probability predictions for data values (x_1, \dots, x_k) not in the sample. The second gives precisely the *outcome prediction* values in (17) above [where in the present case the values are 1 = “a_Rainy” and 0 = “b_Dry”].
- (iii) So to construct the corresponding *prediction success*, one can add a new column, labeled **Success** in the file **Rain_Logit.jmp**, and calculate the success value by the formula for this column, i.e., by

$$\mathbf{If} \left(\begin{array}{ll} \text{Rained_New} == \text{MostLikely Rained_New} & \Rightarrow 1 \\ \text{else} & \Rightarrow 0 \end{array} \right)$$

- (iv) Finally, the *prediction success rate* can be computed as in the column **Success Rate** by the formula

$$\mathbf{ColMean}(\text{Success})$$

- (v) This value shows a 90% success rate, which is quite respectable. (Remember however that this success rate is *only* for the data used to fit the model, and offers no guarantees for future predictions.)

Now that you are familiar with Logistic Regression, it is important to emphasize you can also do STEPWISE Logistic Regression by simply choosing the option “Stepwise” rather than “Nominal Logistic”. The stepwise procedure here is identical with linear regression, and requires no additional discussion.

Finally it should be noted that you can also analyze possible specification errors by letting **Rain** be an indicator-variable version of **Rained** with **Rain = 1** if and only if **Rained = Rainy**, and then defining appropriate “residuals” by **Rain – Prob[Rained]**. By plotting each of your explanatory variables against these residuals, you can look for possible structure in these patterns. For example, if the residuals are U-shaped (i.e., largest at each end of the plot) for variable x , then perhaps a quadratic specification of x would yield a better fit.