Background Note of the Basics of Statistics used in SPSS (*compiled from multiple sources for restricted circulation)*

Compiled by Prof. Koustab Ghosh, IIM Rohtak

Explanatory Note by the Compiler:

The following write-up gives a fair idea about t test, f test, simple linear regression, and basics of logistic regression. This is a conceptual note that needs to be consulted for performing some basic statistical operations using SPSS. SPSS as a software performs all calculations and we need to draw support from these basic concepts in two main ways. First, to decide which test to be used when, and second, to draw inferences from the output obtained in SPSS for interpreting results and taking decisions.

A. Basic Information on the t-Test

Hypothesis: The hypothesis is a tentative explanation based on observations you have made. Your observations may have been followed up with a search of the literature for more information before you develop your hypothesis.

Example: Men's hands are larger than women's hands OR adding fertilizer to a plant makes it grow better.

Null hypothesis: The actual null hypothesis is a more formal statement of your original hypothesis. The null hypothesis is usually written in the following form: There is no significant difference between population A and population B.

Example: There is no significant difference in hand size between males and females. OR There is no significant difference in the growth of fertilized plants vs. unfertilized plants.

The reason we write it in this form is that scientists are basically skeptics and their goal is to prove a hypothesis false. In fact, you can never really prove that a hypothesis is true. In addition, the null hypothesis is used because it allows you to relate your calculations of the difference between the sample means to a standard of zero.

The t-Test: We use this statistical test to compare our sample populations and determine if there is a significant difference between their means. The result of the t-test is a 't' value; this value is then used to determine the p-value (see below).

If we cannot use a statistical test (doesn't have to be a t-test) to determine whether a significant difference exists, then it becomes difficult to convince other scientists that your research is worth anything.

P-value: The p-value is the probability that 't' falls into a certain range. In other words this is the value you use to determine if the difference between the means in your sample populations is significant. For our purposes, a p-value ≤ 0.05 suggests a significant

difference between the means of our sample population and we would reject our null hypothesis. A p-value > 0.05 suggests no significant difference between the means of our sample populations and we would not reject our null hypothesis.

Types of t-tests: There are two types of t-tests, the unpaired and paired t-test that we will use in this course.

Unpaired t-test: This type of t-test is used when you have independent samples. In other words your samples are not directly related to one another. Ex.: Index finger length between males and females.

Paired t-test: In this t-test your samples are related. You collected data before and after some manipulation of your subjects. Ex.: Pulse before and after 3 cups of coffee.

B. Basic Information on the f-Test

The term F-test is based on the fact that these tests use the F-statistic to test the hypotheses. An F-statistic is the ratio of two variances and it was named after Sir Ronald Fisher. Variances measure the dispersal of the data points around the mean. Higher variances occur when the individual data points tend to fall further from the mean.

It's difficult to interpret variances directly because they are in squared units of the data. If you take the square root of the variance, you obtain the standard deviation, which is easier to interpret because it uses the data units. While variances are hard to interpret directly, some statistical tests use them in their equations.

An F-statistic is the ratio of two variances, or technically, two mean squares. Mean squares are simply variances that account for the degrees of freedom (DF) used to estimate the variance.

Think of it this way. Variances are the sum of the squared deviations from the mean. If you have a bigger sample, there are more squared deviations to add up. The result is that the sum becomes larger and larger as you add in more observations. By incorporating the DF, mean squares account for the differing numbers of measurements for each estimate of the variance. Otherwise, the variances are not comparable, and the ratio for the F-statistic is meaningless.

Given that F-tests evaluate the ratio of two variances, you might think it's only suitable for determining whether the variances are equal. Actually, it can do that and a lot more! F-tests are surprisingly flexible because you can include different variances in the ratio to test a wide variety of properties. F-tests can compare the fits of different models, test the overall significance in regression models, test specific terms in linear models, and determine whether a set of means are all equal.

Earlier in this section you saw how to perform a *t*-test to compare a sample mean to an accepted value, or to compare two sample means. In this section, you will see how to use the *F*-test to compare two variances or standard deviations.

When using the *F*-test, you again require a hypothesis, but this time, it is to compare standard deviations. That is, you will test the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ against an appropriate alternate hypothesis.

You calculate the *F*-value as the ratio of the two variances:

where $s_1^2 \geq s_2^2$, so that $F \geq 1$. The degrees of freedom for the numerator and denominator are n_1-1 and n_2-1, respectively. As with the *t*-test, you compare F_{calc} to a tabulated value F_{tab}, to see if you should accept or reject the null hypothesis. As well, you can perform 1- or 2-tailed *F*-tests. The following two examples illustrate the use of 1- and 2-tailed tests.

Example 1

As an example, assume we want to see if a method (Method 1) for measuring the arsenic concentration in soil is significantly more precise than a second method (Method 2). Each method was tested ten times, with, yielding the following values:

Method	Mean (ppm)	Standard Deviation (ppm)
1	6.7	0.8
2	8.2	1.2

A method is more precise if its standard deviation is lower than that of the other method. So we want to test the null hypothesis H_0: $\sigma_2^2 = \sigma_1^2$, against the alternate hypothesis H_A: $\sigma_2^2 > \sigma_1^2$.

Since $s_2 > s_1$, $F_{calc} = s_2^2/s_1^2 = 1.2^2/0.8^2 = 2.25$. The tabulated value for d.o.f. $\nu = 9$ in each case, and a 1-tailed, 95% confidence level is $F_{9,9} = 3.179$. In this case, $F_{calc} < F_{9,9}$, so we accept the null hypothesis that the two standard deviations are equal, and we are 95% confident that any difference in the sample standard deviations is due to random error. We use a 1-tailed test in this case because the only information we are interested in is whether Method 1 is more precise than Method 2.

Example 2

If we are not interested in whether one method is better compared to another, but were simply trying to determine if the variances of were the same or different, we would need to use a 2-tailed test. For instance, assume we made two sets of measurements of ethanol concentration in a sample of vodka using the same instrument, but on two different days. On the first day, we found a standard deviation of $s_1 = 9$ ppm and on the next day we found $s_2 = 2$ ppm. Both datasets comprised 6 measurements. We want to know if we can combine the two datasets, or if there is a significant difference between the datasets, and that we should discard one of them.

As usual, we begin by defining the null hypothesis, H_0: $\sigma_1^2 = \sigma_2^2$, and the alternate hypothesis, H_A: $\sigma_1^2 \neq \sigma_2^2$. The "≠" sign indicates that this is a 2-tailed test, because we are

interested in both cases: $\sigma_1^2 > \sigma_2^2$ and $\sigma_1^2 < \sigma_2^2$. For the *F*-test, you can perform a 2-tailed test by multiplying the confidence level P by 2, so from a table for a 1-tailed test at the P = 0.05 confidence level, we would perform a 2-tailed test at P = 0.10, or a 90% confidence level.

For this dataset, $s_2 > s_1$, $F_{calc} = s_1^2 / s_2^2 = 9^2/2^2 = 20.25$. The tabulated value for $v = 5$ at 90% confidence is $F_{5,5} = 5.050$. Since $F_{calc} > F_{5,5}$, we reject the null hypothesis, and can say with 90% certainty that there is a difference between the standard deviations of the two methods.

Tables for other confidence levels can be found in most statistics or analytical chemistry textbooks. Be careful when using these tables, to pay attention to whether the table is for a 1- or a 2-tailed test. In most cases, tables are given for 2-tailed tests, so you can divide by 2 for the 1-tailed test. For the *F*-test, always ensure that the larger standard deviation is in the numerator, so that $F \geq 1$.

C. Basics of Simple Linear Regression

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable*and is referred to as Y. The variable we are basing our predictions on is called the *predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

The example data in Table 1 are plotted in Figure 1. You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

Table 1. Example data.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

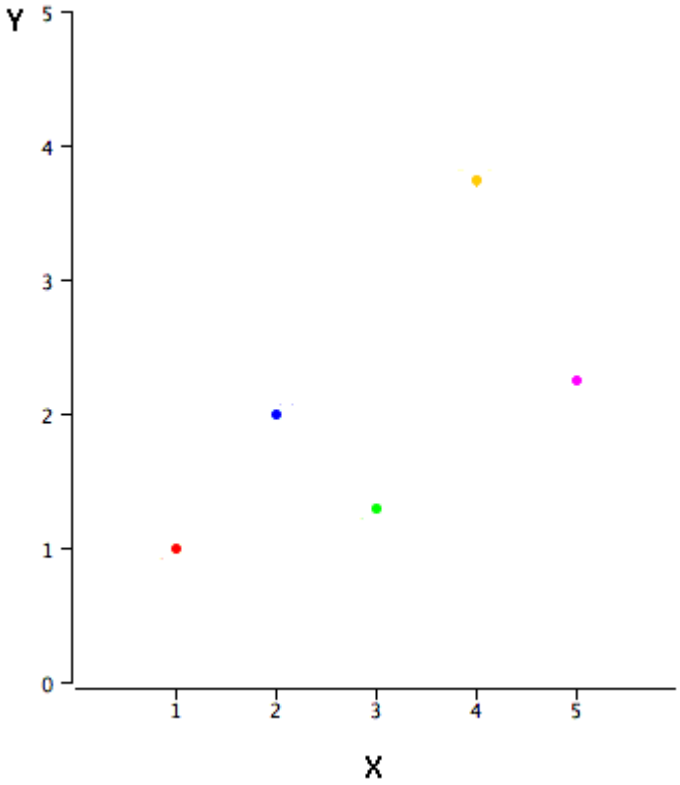

Figure 1. A scatter plot of the example data.

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a *regression line*. The black diagonal line in Figure 2 is the regression line and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.

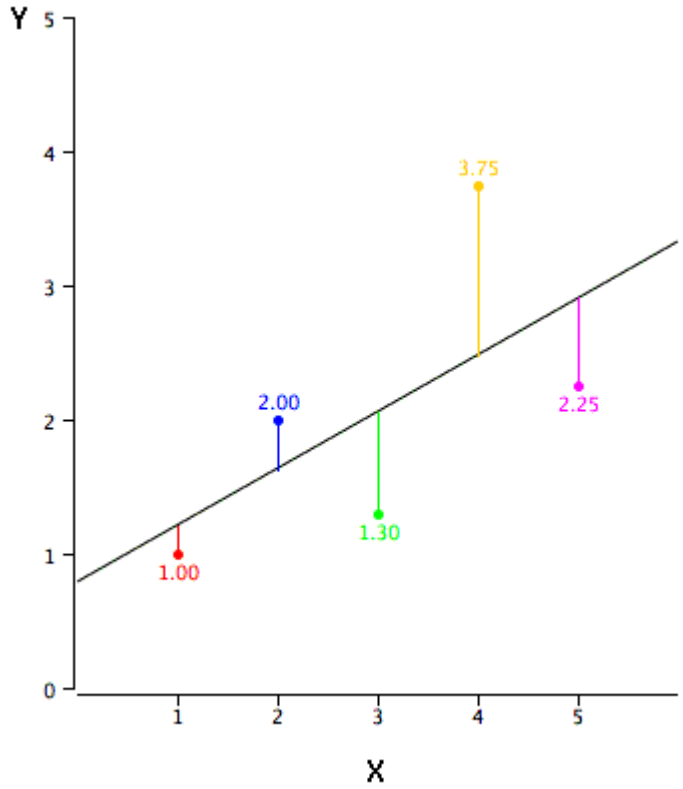

Figure 2. A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

The error of prediction for a point is the value of the point minus the predicted value (the value on the line). Table 2 shows the predicted values (Y') and the errors of prediction (Y-Y'). For example, the first point has a Y of 1.00 and a predicted Y (called Y') of 1.21. Therefore, its error of prediction is -0.21.

Table 2. Example data.

X	Y	Y'	Y-Y'	$(Y-Y')^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

You may have noticed that we did not specify what is meant by "best-fitting line." By far, the most commonly-used criterion for the best-fitting line is the line that minimizes the sum of the squared errors of prediction. That is the criterion that was used to find the line in Figure 2. The last column in Table 2 shows the squared errors of prediction. The sum of the squared errors of prediction shown in Table 2 is lower than it would be for any other regression line.

The formula for a regression line is

$$Y' = bX + A$$

where Y' is the predicted score, b is the slope of the line, and A is the Y intercept. The equation for the line in Figure 2 is

$$Y' = 0.425X + 0.785$$

For X = 1,

$$Y' = (0.425)(1) + 0.785 = 1.21.$$

For X = 2,

$$Y' = (0.425)(2) + 0.785 = 1.64.$$

COMPUTING THE REGRESSION LINE

In the age of computers, the regression line is typically computed with statistical software. However, the calculations are relatively easy, and are given here for anyone who is interested. The calculations are based on the statistics shown in Table 3. M_X is the mean of X, M_Y is the mean of Y, s_X is the standard deviation of X, s_Y is the *standard deviation* of Y, and r is the *correlation* between X and Y.

Table 3. Statistics for computing the regression line.

M_X	M_Y	s_X	s_Y	r
3	2.06	1.581	1.072	0.627

The slope (b) can be calculated as follows:

$$b = r\, s_Y/s_X$$

and the intercept (A) can be calculated as

$$A = M_Y - bM_X.$$

For these data,

$$b = (0.627)(1.072)/1.581 = 0.425$$

$$A = 2.06 - (0.425)(3) = 0.785$$

Note that the calculations have all been shown in terms of sample statistics rather than population parameters. The formulas are the same; simply use the parameter values for means, standard deviations, and the correlation.

STANDARDIZED VARIABLES

The regression equation is simpler if variables are *standardized* so that their means are equal to 0 and standard deviations are equal to 1, for then b = r and A = 0. This makes the regression line:

$$Z_{Y'} = (r)(Z_X)$$

where $Z_{Y'}$ is the predicted standard score for Y, r is the correlation, and Z_X is the standardized score for X. Note that the slope of the regression equation for standardized variables is r.

A REAL EXAMPLE

The case study "SAT and College GPA" contains high school and university grades for 105 computer science majors at a local state school. We now consider how we could predict a student's university GPA if we knew his or her high school GPA.

Figure 3 shows a scatter plot of University GPA as a function of High School GPA. You can see from the figure that there is a strong positive relationship. The correlation is 0.78. The regression equation is

University GPA' = (0.675)(High School GPA) + 1.097

Therefore, a student with a high school GPA of 3 would be predicted to have a university GPA of

University GPA' = (0.675)(3) + 1.097 = 3.12.

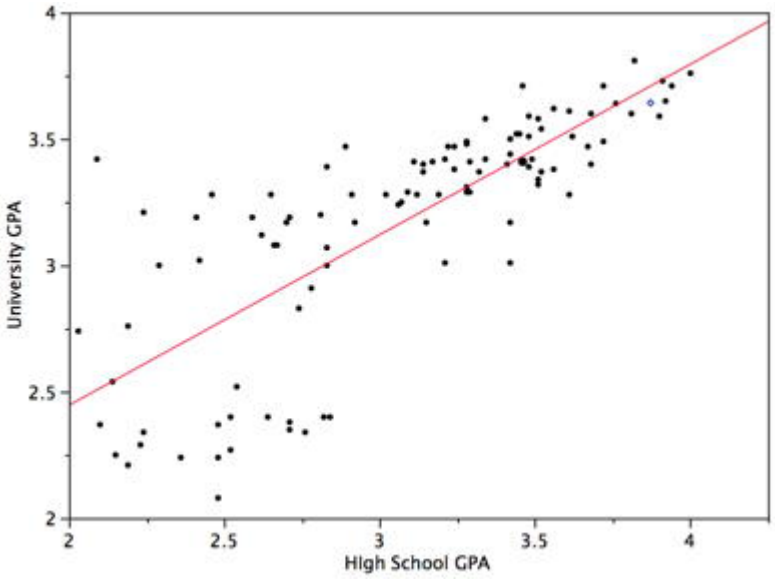

Figure 3. University GPA as a function of High School GPA.

D. Basics of Logistic Regression

Why use logistic regression?

There are many important research topics for which the dependent variable is "limited" (discrete not continuous). Researchers often want to analyze whether some event occurred or not, such as voting, participation in a public program, business success or failure, morbidity, mortality, a hurricane and etc.

Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable (coded 0, 1).

A data set appropriate for logistic regression might look like this:

Descriptive Statistics					
Variable	**N**	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
YES	122	.00	1.00	.6393	.4822
BAG	122	.00	7.00	1.5082	1.8464
COST	122	9.00	953.00	416.5492	285.4320
INCOME	122	5000.00	85000.00	38073.7705	18463.1274
Valid N (listwise)	122				

*This data is from a U.S. Department of the Interior survey (conducted by U.S. Bureau of the Census) which looks at a yes/no response to a question about the "willingness to pay" higher travel costs for deer hunting trips in North Carolina (a more complete description of this data can be found here).

The linear probability model

"Why shouldn't I just use ordinary least squares?" Good question.

Consider the linear probability (LP) model:

$Y = \boldsymbol{a} + \boldsymbol{B}X + e$

where

- Y is a dummy dependent variable, =1 if event happens, =0 if event doesn't happen,
- $\boldsymbol{a}$ is the coefficient on the constant term,
- $\boldsymbol{B}$ is the coefficient(s) on the independent variable(s),
- X is the independent variable(s), and
- e is the error term.

Use of the LP model generally gives you the correct answers in terms of the sign and significance level of the coefficients. The predicted probabilities from the model are usually where we run into trouble. There are 3 problems with using the LP model:

1. The error terms are heteroskedastic (*heteroskedasticity occurs when the variance of the dependent variable is different with different values of the independent variables*): var(e)= p(1-p), where p is the probability that EVENT=1. Since P depends on X the "classical regression assumption" that the error term does not depend on the Xs is violated.
2. e is not normally distributed because P takes on only two values, violating another "classical regression assumption"
3. The predicted probabilities can be greater than 1 or less than 0 which *can be a problem* if the predicted values are used in a subsequent analysis. Some people try to solve this problem by setting probabilities that are greater than (less than) 1 (0) to be equal to 1 (0). This amounts to an interpretation that a high probability of the Event (Nonevent) occuring is considered a sure thing.

<u>The logistic regression model</u>

The "logit" model solves these problems:

$\ln[p/(1-p)] = \boldsymbol{a} + \boldsymbol{B}X + e$ or

$[p/(1-p)] = \exp(\boldsymbol{a} + \boldsymbol{B}X + e)$

where:

- ln is the natural logarithm, $\log_{exp}$, where exp=2.71828…
- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"
- all other components of the model are the same.

The logistic regression model is simply a non-linear transformation of the linear regression. The "logistic" distribution is an S-shaped distribution function which is similar to the standard-normal distribution (which results in a probit regression model) but easier to work with in most applications (the probabilities are easier to calculate). The logit distribution constrains the estimated probabilities to lie between 0 and 1.

For instance, the estimated probability is:

$p = 1/[1 + \exp(-\boldsymbol{a} - \boldsymbol{B}X)]$

With this functional form:

- if you let $\boldsymbol{a} + \boldsymbol{B}X = 0$, then p = .50
- as $\boldsymbol{a} + \boldsymbol{B}X$ gets really big, p approaches 1
- as $\boldsymbol{a} + \boldsymbol{B}X$ gets really small, p approaches 0.

A graphical comparison of the linear probability and logistic regression models is illustrated here.

Interpreting logit coefficients

The estimated coefficients must be interpreted with care. Instead of the slope coefficients (***B***) being the rate of change in Y (the dependent variables) as X changes (as in the LP model or OLS regression), now the slope coefficient is interpreted as the rate of change in the "log odds" as X changes. This explanation is not very intuitive. It is possible to compute the more intuitive "marginal effect" of a continuous independent variable on the probability. The marginal effect is

$dp/d\mathbf{B} = f(\mathbf{B}X)\mathbf{B}$

where *f*(.) is the density function of the cumulative probability distribution function [F(**B**X), which ranges from 0 to 1]. The marginal effects depend on the values of the independent variables, so, it is often useful to evaluate the marginal effects at the means of the independent variables. (*SPSS doesn't have an option for the marginal effects. If you need to compute marginal effects you can use the LIMDEP statistical package which is available on the academic mainframe*.)

An interpretation of the logit coefficient which is usually more intuitive (especially for dummy independent variables) is the "odds ratio"-- exp***B*** is the effect of the independent variable on the "odds ratio" [the odds ratio is the probability of the event divided by the probability of the nonevent]. For example, if $\exp\boldsymbol{B}_3 = 2$, then a one unit change in X_3 would make the event twice as likely (.67/.33) to occur. Odds ratios equal to 1 mean that there is a 50/50 chance that the event will occur with a small change in the independent variable. Negative coefficients lead to odds ratios less than one: if $\exp\boldsymbol{B}_2 = .67$, then a one unit change in X_2 leads to the event being less likely (.40/.60) to occur. {Odds ratios less than 1 (negative coefficients) tend to be harder to interpret than odds ratios greater than one(positive coefficients).} Note that odds ratios for continuous independent variables tend to be close to one, this does NOT suggest that the coefficients are insignificant. Use the Wald statistic (see below) to test for statistical significance.

Estimation by maximum likelihood

[For those of you who just NEED to know ...] Maximum likelihood estimation (MLE) is a statistical method for estimating the coefficients of a model. MLE is usually used as an alternative to non-linear least squares for nonlinear equations.

The likelihood function (L) measures the probability of observing the particular set of dependent variable values (p_1, p_2, ..., p_n) that occur in the sample. It is written as the probability of the product of the dependent variables:

$L = \text{Prob}\ (p_1 * p_2 * * * p_n)$

The higher the likelihood function, the higher the probability of observing the ps in the sample. MLE involves finding the coeffients (***a***, ***B***) that makes the log of the likelihood function (LL < 0) as large as possible or -2 times the log of the likelihood function (-2LL) as small as possible. The maximum likelihood estimates solve the following condition:

$\{Y - p(Y=1)\}X_i = 0$, summed over all observations

{*or something like that ...* }

Hypothesis testing

Testing the hypothesis that a coefficient on an independent variable is significantly different from zero is similar to OLS models. The Wald statisitic for the ***B*** coefficient is:

$\text{Wald} = [\boldsymbol{B}/\text{s.e.}_{\boldsymbol{B}}]^2$

which is distributed chi-square with 1 degree of freedom. The Wald is simply the square of the (asymptotic) t-statistic.

The probability of a YES response from the data above was estimated with the logistic regression procedure in SPSS (click on "statistics," "regression," and "logistic"). The SPSS results look like this:

Variables in the Equation							
Variable	**B**	**S.E.**	**Wald**	**df**	**Sig**	**R**	**Exp(B)**
	[1]	[2]	[3]		[4]	[5]	[6]
BAG	0.2639	0.1239	4.5347	1	0.0332	0.1261	1.302
INCOME	4.63E-07	1.07E-05	0.0019	1	0.9656	0	1
COST	-0.0018	0.0007	6.5254	1	0.0106	-0.1684	0.9982
Constant	0.9691	0.569	2.9005	1	0.0885		
Notes:							
[1] B is the estimated logit coefficient							
[2] S.E. is the standard error of the coefficient							
[3] Wald = $[B/S.E.]^2$							
[4] "Sig" is the significance level of the coefficient: "the coefficient on BAG is significant at the .03 (97% confidence) level."							
[5] The "Partial R" = sqrt{[(Wald-2)/(-2*LL(a)]}; see below for LL(a)							
[6] Exp(B) is the "odds ratio" of the individual coefficient.							

Evaluating the overall performance of the model

There are several statistics which can be used for comparing alternative models or evaluating the performance of a single model:

1. The model likelihood ratio (LR), or chi-square, statistic is

$LR[i] = -2[LL(\boldsymbol{a}) - LL(\boldsymbol{a},\boldsymbol{B})]$

or as you are reading SPSS printout:

LR[i] = [-2 Log Likelihood (of beginning model)]

- [-2 Log Likelihood (of ending model)].

where the model LR statistic is distributed chi-square with i degrees of freedom, where i is the number of independent variables. The "unconstrained model", LL(**a**,$\mathbf{B}_i$), is the log-likelihood function evaluated with all independent variables included and the "constrained model" is the log-likelihood function evaluated with only the constant included, LL(**a**).

Use the Model Chi-Square statistic to determine if the overall model is statistically significant.

2. The "Percent Correct Predictions" statistic assumes that if the estimated p is greater than or equal to .5 then the event is expected to occur and not occur otherwise. By assigning these probabilities 0s and 1s the following table is constructed:

Classification Table for YES The Cut Value is .50				
		Predicted		% Correct
		0	1	
Observed	0	9	35	20.25%
	1	4	74	94.87%
Overall				68.03%

the bigger the % Correct Predictions, the better the model.

3. Most OLS researchers like the R^2 statistic. It is the proportion of the variance in the dependent variable which is explained by the variance in the independent variables. There is NO equivalent measure in logistic regression. However, there are several "Pseudo" R^2 statistics. One psuedo R^2 is the McFadden's-R^2 statistic (sometimes called the likelihood ratio index [LRI]):

McFadden's-R^2= 1 - [LL(***a***,***B***)/LL(***a***)]
= 1 - [-2LL(***a***,***B***)/-2LL(***a***)]

where the R^2 is a scalar measure which varies between 0 and (somewhat close to) 1 much like the R^2 in a LP model. Expect your Pseudo R^2s to be much less than what you would expect in LP model, however. Because the LRI depends on the ratio of the beginning and ending log-likelihood functions, it is very difficult to "maximize the R^2" in logistic regression.

The Pseudo-R^2 in logistic regression is best used to compare different specifications of the same model. Don't try to compare models with different data sets with the Pseudo-R^2 [referees will yell at you ...].

Other Pseudo-R^2 statistics are printed in SPSS output but [YIKES!] I can't figure out how these are calculated (even after consulting the manual and the SPSS discussion list)!?!

Source: SPSS Output		
(-2)*Initial LL	[1]	159.526

(-2)*Ending LL	[2]	147.495	
Goodness of Fit	[3]	123.18	
Cox & Snell-R^2		0.094	
Nagelkerke-R^2		0.129	
	Chi-Square [4]	df	Significance
Model	12.031	3	0.0073
Notes:			
[1] LL(a) = 159.526/(-2) = -79.763			
[2] LL(a,B) = 147.495/(-2) = -73.748			
[3] GF = [Y - P(Y=1)]^2/[Y - P(Y=1)]			
[4] Chi-Square = -2[LL(a)-LL(a,B)] = 159.526 - 147.495			
McFadden's-R^2 = 1- (147.495/159.526) = 0.075			