Chapter 16

Analyzing Experiments with Categorical Outcomes

Analyzing data with non-quantitative outcomes

All of the analyses discussed up to this point assume a Normal distribution for the outcome (or for a transformed version of the outcome) at each combination of levels of the explanatory variable(s). This means that we have only been covering statistical methods appropriate for quantitative outcomes. It is important to realize that this restriction only applies to the outcome variable and not to the explanatory variables. In this chapter statistical methods appropriate for categorical outcomes are presented.

16.1 Contingency tables and chi-square analysis

This section discusses analysis of experiments or observational studies with a categorical outcome and a single categorical explanatory variable. We have already discussed methods for analysis of data with a quantitative outcome and categorical explanatory variable(s) (ANOVA and ANCOVA). The methods in this section are also useful for observational data with two categorical "outcomes" and no explanatory variable.

16.1.1   Why ANOVA and regression don't work

There is nothing in most statistical computer programs that would prevent you from analyzing data with, say, a two-level categorical outcome (usually designated generically as "success" and "failure") using ANOVA or regression or ANCOVA. But if you do, your conclusion will be wrong in a number of different ways. The basic reason that these methods don't work is that the assumptions of Normality and equal variance are strongly violated. Remember that these assumptions relate to groups of subjects with the same levels of all of the explanatory variables. The Normality assumption says that in each of these groups the outcomes are Normally distributed. We call ANOVA, ANCOVA, and regression "robust" to this assumption because moderate deviations from Normality alter the null sampling distributions of the statistics from which we calculate p-values only a small amount. But in the case of a categorical outcome with only a few (as few as two) possible outcome values, the outcome is so far from the smooth bell-shaped curve of a Normal distribution, that the null sampling distribution is drastically altered and the p-value completely unreliable.

The equal variance assumption is that, for any two groups of subjects with different levels of the explanatory variables between groups and the same levels within groups, we should find that the variance of the outcome is the same. If we consider the case of a binary outcome with coding 0=failure and 1=success, the variance of the outcome can be shown to be equal to $p_i(1 - p_i)$ where p_i is the probability of getting a success in group i (or, equivalently, the mean outcome for group i). Therefore groups with different means have different variances, violating the equal variance assumption.

A second reason that regression and ANCOVA are unsuitable for categorical outcomes is that they are based on the prediction equation $E(Y) = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k$, which both is inherently quantitative, and can give numbers out of range of the category codes. The least unreasonable case is when the categorical outcome is ordinal with many possible values, e.g., coded 1 to 10. Then for any particular explanatory variable, say, β_i, a one-unit increase in x_i is associated with a β_i unit change in outcome. This works only over a limited range of x_i values, and then predictions are outside the range of the outcome values.

For binary outcomes where the coding is 0=failure and 1=success, a mean outcome of, say, 0.75 corresponds to 75% successes and 25% failures, so we can think of the prediction as being the probability of success. But again, outside of some limited range of x_i values, the predictions will correspond to the absurdity

of probabilities less than 0 or greater than 1.

And for nominal categorical variables with more than two levels, the prediction is totally arbitrary and meaningless.

> **Using statistical methods designed for Normal, quantitative outcomes when the outcomes are really categorical gives wrong p-values due to violation of the Normality and equal variance assumptions, and also gives meaningless out-of-range predictions for some levels of the explanatory variables.**

16.2 Testing independence in contingency tables

16.2.1 Contingency and independence

A contingency table counts the number of cases (subjects) for each combination of levels of two or more categorical variables. An equivalent term is cross-tabulation (see Section 4.4.1). Among the definitions for "contingent" in the The Oxford English Dictionary is "Dependent for its occurrence or character on or upon some prior occurrence or condition". Most commonly when we have two categorical measures on each unit of study, we are interested in the question of whether the probability distribution (see section 3.2) of the levels of one measure depends on the level of the other measure, or if it is independent of the level of the second measure. For example, if we have three treatments for a disease as one variable, and two outcomes (cured and not cured) as the other outcome, then we are interested in the probabilities of these two outcomes for each treatment, and we want to know if the observed data are consistent with a null hypothesis that the true underlying probability of a cure is the same for all three treatments.

In the case of a clear identification of one variable as explanatory and the other as outcome, we focus on the probability distribution of the outcome and how it changes or does not change when we look separately at each level of the explanatory variable. The "no change" case is called independence, and indicates that knowing the level of the (purported) explanatory variable tells us no more about the possible outcomes than ignoring or not knowing it. In other words, if the

variables are independent, then the "explanatory" variable doesn't really explain anything. But if we find evidence to reject the null hypothesis of independence, then we do have a true explanatory variable, and knowing its value allows us to refine our predictions about the level of the other variable.

Even if both variables are outcomes, we can test their association in the same way as just mentioned. In fact, the conclusions are always the same when the roles of the explanatory and outcome variables are reversed, so for this type of analysis, choosing which variable is outcome vs. explanatory is immaterial.

Note that if the outcome has only two possibilities then we only need the probability of one level of the variable rather than the full probability distribution (list of possible values and their probabilities) for each level of the explanatory variable. Of course, this is true simply because the probabilities of all levels must add to 100%, and we can find the other probability by subtraction.

The usual statistical test in the case of a categorical outcome and a categorical explanatory variable is whether or not the two variables are independent, which is equivalent to saying that the probability distribution of one variable is the same for each level of the other variable.

16.2.2   Contingency tables

It is a common situation to measure two categorical variables, say X (with k levels) and Y (with m levels) on each subject in a study. For example, if we measure gender and eye color, then we record the level of the gender variable and the level of the eye color variable for each subject. Usually the first task after collecting the data is to present it in an understandable form such as a **contingency table** (also known as a cross-tabulation).

For two measurements, one with k levels and the other with m levels, the contingency table is a $k \times m$ table with cells for each combination of one level from each variable, and each cell is filled with the corresponding count (also called **frequency**) of units that have that pair of levels for the two categorical variables.

For example, table 16.1 is a (fake) contingency table showing the results of asking 271 college students what their favorite music is and what their favorite ice

		favorite ice cream				
		chocolate	vanilla	strawberry	other	total
favorite music	rap	5	10	7	38	60
	jazz	8	9	23	6	46
	classical	12	3	4	3	22
	rock	39	10	15	9	73
	folk	10	22	8	8	48
	other	4	7	5	6	22
	total	78	61	62	70	271

Table 16.1: Basic ice cream and music contingency table.

cream flavor is. This table was created in SPSS by using the Cross-tabs menu item under Analysis / Descriptive Statistics. In this simple form of a contingency table we see the **cell counts** and the **marginal counts**. The margins are the extra column on the right and the extra row at the bottom. The cells are the rest of the numbers in the table. Each cell tells us how many subjects gave a particular pair of answers to the two questions. For example, 23 students said both that strawberry is their favorite ice cream flavor and that jazz is their favorite type of music. The right margin sums over ice cream types to show that, e.g., a total of 60 students say that rap is their favorite music type. The bottom margin sums over music types to show that, e.g,, 70 students report that their favorite flavor of ice cream is neither chocolate, vanilla, nor strawberry. The total of either margin, 271, is sometimes called the "grand total" and represent the total number of subjects.

We can also see, from the margins, that rock is the best liked music genre, and classical is least liked, though there is an important degree of arbitrariness in this conclusion because the experimenter was free to choose which genres were in or not in the "other" group. (The best practice is to allow a "fill-in" if someone's choice is not listed, and then to be sure that the "other" group has no choices with larger frequencies that any of the explicit non-other categories.) Similarly, chocolate is the most liked ice cream flavor, and subject to the concern about defining "other", vanilla and strawberry are nearly tied for second.

Before continuing to discuss the form and content of contingency tables, it is good to stop and realize that the information in a contingency table represents results from a sample, and other samples would give somewhat different results. As usual, any differences that we see in the sample may or may not reflect real

		favorite ice cream				
		chocolate	vanilla	strawberry	other	total
favorite music	rap	5 8.3%	10 17.7%	7 11.7%	38 63.3%	60 100%
	jazz	8 17.4%	9 19.6%	23 50.0%	6 13.0%	46 100%
	classical	12 54.5%	3 13.6%	4 18.2%	3 13.6%	22 100%
	rock	39 53.4%	10 13.7%	15 20.5%	9 12.3%	73 100%
	folk	10 20.8%	22 45.8%	8 16.7%	8 16.7%	48 100%
	other	4 18.2%	7 31.8%	5 22.7%	6 27.3%	22 100%
	total	78 28.8%	61 22.5%	62 22.9%	70 25.8%	271 100%

Table 16.2: Basic ice cream and music contingency table with row percents.

differences in the population, so you should be careful not to over-interpret the information in the contingency table. In this sense it is best to think of the contingency table as a form of EDA. We will need formal statistical analysis to test hypotheses about the population based on the information in our sample.

Other information that may be present in a contingency table includes various percentages. So-called **row percents** add to 100% (in the right margin) for each row of the table, and **column percents** add to 100% (in the bottom margin) for each column of the table.

For example, table 16.2 shows the ice cream and music data with row percents. In SPSS the Cell button brings up check boxes for adding row and/or column percents. If one variable is clearly an outcome variable, then the most useful and readable version of the table is the one with cell counts plus percentages that add up to 100% across all levels of the outcome for each level of the explanatory variable. This makes it easy to compare the outcome distribution across levels of the explanatory variable. In this example there is no clear distinction of the roles of the two measurements, so arbitrarily picking one to sum to 100% is a good approach.

Many important things can be observed from this table. First, we should look for the 100% numbers to see which way the percents go. Here we see 100% on the right side of each row. So for any music type we can see the frequency of each flavor answer and those frequencies add up to 100%. We should think of those row percents as estimates of the true population probabilities of the flavors for each given music type.

Looking at the bottom (marginal) row, we know that, e.g., averaging over all music types, approximately 26% of students like "other" flavors best, and approximately 29% like chocolate best. Of course, if we repeat the study, we would get somewhat different results because each study looks at a different random sample from the population of interest.

In terms of the main hypothesis of interest, which is whether or not the two questions are independent of each other, it is equivalent to ask whether all of the row probabilities are similar to each other and to the marginal row probabilities. Although we will use statistical methods to assess independence, it is worthwhile to examine the row (or column) percentages for equality. In this table, we see rather large differences, e.g., chocolate is high for classical and rock music fans, but low for rap music fans, suggesting lack of independence.

A contingency table summarizes the data from an experiment or observational study with two or more categorical variables. Comparing a set of marginal percentages to the corresponding row or column percentages at each level of one variable is good EDA for checking independence.

16.2.3   Chi-square test of Independence

The most commonly used test of independence for the data in a contingency table is the **chi-square test of independence**. In this test the data from a k by m contingency table are reduced to a single statistic usually called either X^2 or χ^2 (chi-squared), although X^2 is better because statistics usually have Latin, not Greek letters. The null hypothesis is that the two categorical variables are independent, or equivalently that the distribution of either variable is the same at each level of the other variable. The alternative hypothesis is that the two variables are

not independent, or equivalently that the distribution of one variable depends on (varies with) the level of the other.

If the null hypothesis of independence is true, then the X^2 statistic is **asymptotically distributed** as a chi-square distribution (see section 3.9.6) with $(k-1)(m-1)$ df. Under the alternative hypothesis of non-independence the X^2 statistic will be larger on average. The p-value is the area under the null sampling distribution larger than the observed X^2 statistic. The term asymptotically distributed indicates that the null sampling distribution can not be computed exactly for a small sample size, but as the sample size increases, the null sampling distribution approaches the shape of a particular known distribution, which is the chi-square distribution in the case of the X^2 statistic. So the p-values are reliable for "large" sample sizes, but not for small sample sizes. Most textbooks quote a rule that no cell of the expected counts table (see below) can have less than five counts for the X^2 test to be reliable. This rule is conservative, and somewhat smaller counts also give reliable p-values.

Several alternative statistics are sometimes used instead of the chi-square statistic (e.g., likelihood ratio statistic or Fisher exact test), but these will not be covered here. It is important to realize that these various tests may disagree for small sample sizes and it is not clear (or meaningful to ask) which one is "correct".

The calculation of the X^2 statistic is based on the formula

$$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(\text{Observed}_{ij} - \text{Expected}_{ij})^2}{\text{Expected}_{ij}}$$

where k and m are the number of rows and columns in the contingency table (i.e., the number of levels of the categorical variables), Observed_{ij} is the observed count for the cell with one variable at level i and the other at level j, and Expected_{ij} is the expected count based on independence. The basic idea here is that each cell contributes a non-negative amount to the sum, that a cell with an observed count very different from expected contributes a lot, and that "a lot" is relative to the expected count (denominator).

Although a computer program is ordinarily used for the calculation, an understanding of the principles is worthwhile. An "expected counts" table can be constructed by looking at either of the marginal percentages, and then computing the expected counts by multiplying each of these percentages by the total counts in the other margin. Table 16.3 shows the expected counts for the ice cream example. For example, using the percents in the bottom margin of table 16.2, if the two

		favorite ice cream				
		chocolate	vanilla	strawberry	other	total
	rap	17.3	13.5	13.7	15.5	60
	jazz	13.2	10.4	10.5	11.9	46
favorite	classical	6.3	5.0	5.0	5.7	22
music	rock	21.0	16.4	16.7	18.9	73
	folk	13.8	10.8	11.0	12.4	48
	other	6.3	5.0	5.0	5.7	22
	total	78	61	62	70	271

Table 16.3: Expected counts for ice cream and music contingency table.

variables are independent, then we expect 22.9% of people to like strawberry best among each group of people defined by their favorite music. Because 73 people like rock best, under the null hypothesis of independence, we expect (on average) $0.229 * 73 = 16.7$ people to like rock and strawberry best, as shown in table 16.3. Note that there is no reason that the expected counts should be whole numbers, even though observed counts must be.

By combing the observed data of table 16.1 with the expected values of table 16.3, we have the information we need to calculate the X^2 statistic. For the ice cream data we find that

$$X^2 = \left(\frac{(5 - 17.3)^2}{17.3}\right) + \left(\frac{(10 - 13.5)^2}{13.5}\right) + \cdots + \left(\frac{(6 - 5.7)^2}{5.7}\right) = 112.86.$$

So for the ice cream example, rock paired with chocolate shows a big deviation from independence and of the 24 terms of the X^2 sum, that cell contributes $(39 - 21.0)^2/21.0 = 15.3$ to the total of 112.86. There are far more people who like that particular combination than would be expected under independence. To test if all of the deviations are consistent with chance variation around the expected values, we compare the X^2 statistic to the χ^2 distribution with $(6-1)(4-1) = 15$ df. This distribution has 95% of its probability below 25.0, so with $X^2 = 112.86$, we reject H_0 at the usual $\alpha = 0.05$ significance level. In fact, only 0.00001 of the probability is above 50.5, so the p-value is far less than 0.05. We reject the null hypothesis of independence of ice cream and music preferences in favor of the conclusions that the distribution of preference of either variable *does* depend on preference for the other variable.

You can choose among several ways to express violation (or non-violation) of the null hypothesis for a "chi-square test of independence" of two categorical variables. You should use the context of the problem to decide which one best expresses the relationship (or lack of relationship) between the variables. In this problem it is correct to say any of the following: ice cream preference is not independent of music preference, or ice cream preference depends on or differs by music preference, or music preference depends on or differs by ice cream preference, or knowing a person's ice cream preference helps in predicting their music preference, or knowing a person's music preference helps in predicting their ice cream preference.

The chi-square test is based on a statistic that is large when the observed cell counts differ markedly from the expected counts under the null hypothesis condition of independence. The corresponding null sampling distribution is a chi-square distribution if no expected cell counts are too small.

Two additional points are worth mentioning in this abbreviated discussion of testing independence among categorical variables. First, because we want to avoid very small expected cell counts to assure the validity of the chi-square test of independence, it is common practice to combine categories with small counts into combined categories. Of course, this must be done in some way that makes sense in the context of the problem.

Second, when the contingency table is larger than 2 by 2, we need a way to perform the equivalent of contrast tests. One simple solution is to create subtables corresponding to the question of interest, and then to perform a chi-square test of independence on the new table. To avoid a high Type 1 error rate we need to make an adjustment, e.g., by using a Bonferroni correction, if this is post-hoc testing. For example to see if chocolate preference is higher for classical than jazz, we could compute chocolate vs. non-chocolate counts for the two music types to get table 16.4. This gives a X^2 statistic of 8.2 with 1 df, and a p-value of 0.0042. If this is a post-hoc test, we need to consider that there are 15 music pairs and 4 flavors plus 6 flavor pairs and 6 music types giving 4*15+6*6=96 similar tests, that might just as easily have been noticed as "interesting". The Bonferroni correction implies using a new alpha value of 0.05/96=0.00052, so because 0.0042 > 0.00052, we cannot make the post-hoc conclusion that chocolate preference differs for jazz vs. classical. In other words, if the null hypothesis of independence is true, and we

| | | favorite ice cream | | |
		chocolate	not chocolate	total
favorite music	jazz	8	38	46
		17.4%	82.6%	100%
	classical	12	10	22
		54.5%	45.5%	100%
	total	20	48	68
		29.4%	70.6%	100%

Table 16.4: Cross-tabulation of chocolate for jazz vs. classical.

data snoop looking for pairs of categories of one factor being different for presence vs. absence of a particular category of the other factor, finding that one of the 96 different p-values is 0.0042 is not very surprising or unlikely.

16.3 Logistic regression

16.3.1 Introduction

Logistic regression is a flexible method for modeling and testing the relationships between one or more quantitative and/or categorical explanatory variables and one **binary** (i.e., two level) categorical outcome. The two levels of the outcome can represent anything, but generically we label one outcome "success" and the other "failure". Also, conventionally, we use code 1 to represent success and code 0 to represent failure. Then we can look at logistic regression as modeling the success probability as a function of the explanatory variables. Also, for any group of subjects, the 0/1 coding makes it true that the mean of Y represents the observed fraction of successes for that group.

Logistic regression resembles ordinary linear regression in many ways. Besides allowing any combination of quantitative and categorical explanatory variables (with the latter in indicator variable form), it is appropriate to include functions of the explanatory variables such as log(x) when needed, as well as products of pairs of explanatory variables (or more) to represent interactions. In addition, there is usually an intercept parameter (β_0) plus one parameter for each explanatory variable (β_1 through β_k), and these are used in the linear combination form: $\beta_0 +$

$x_1\beta_1 + \cdots + x_k\beta_k$. We will call this sum **eta** (written η) for convenience.

Logistic regression differs from ordinary linear regression because its outcome is binary rather than quantitative. In ordinary linear regression the structural (means) model is that $E(Y) = \eta$. This is inappropriate for logistic regression because, among other reasons, the outcome can only take two arbitrary values, while eta can take any value. The solution to this dilemma is to use the means model

$$\log\left(\frac{E(Y)}{1 - E(Y)}\right) = \log\left(\frac{\Pr(Y = 1)}{\Pr(Y = 0)}\right) = \eta.$$

Because of the 0/1 coding, $E(Y)$, read as the "expected value of Y" is equivalent to the probability of success, and $1 - E(Y)$ is the probability of failure. The ratio of success to failure probabilities is called the odds. Therefore our means model for logistic regression is that the log of the odds (or just "log odds") of success is equal to the linear combination of explanatory variables represented as eta. In other words, for any explanatory variable j, if $\beta_j > 0$ then an increase in that variable is associated with an increase in the chance of success and vice versa.

The means model for logistic regression is that the log odds of success equals a linear combination of the parameters and explanatory variables.

A shortcut term that is often used is **logit** of success, which is equivalent to the log odds of success. With this terminology the means model is logit(S)=η, where S indicates success, i.e., Y=1.

It takes some explaining and practice to get used to working with odds and log odds, but because this form of the means model is most appropriate for modeling the relationship between a set of explanatory variables and a binary categorical outcome, it's worth the effort.

First consider the term **odds**, which will always indicate the odds of success for us. By definition

$$\text{odds}(Y = 1) = \frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)} = \frac{\Pr(Y = 1)}{\Pr(Y = 0)}.$$

The odds of success is defined as the ratio of the probability of success to the probability of failure. The odds of success (where Y=1 indicates success) contains

Pr($Y = 1$)	Pr($Y = 0$)	Odds	Log Odds
0	1	0	$-\infty$
0.1	0.9	1/9	-2.197
0.2	0.8	0.25	-1.383
0.25	0.75	1/3	-1.099
1/3	2/3	0.5	-0.693
0.5	0.5	1	0.000
2/3	1/3	2	0.693
0.75	0.25	3	1.099
0.8	0.2	4	1.386
0.9	0.1	9	2.197
1	0	∞	∞

Table 16.5: Relationship between probability, odds and log odds.

the same information as the probability of success, but is on a different scale. Probability runs from 0 to 1 with 0.5 in the middle. Odds runs from 0 to ∞ with 1.0 in the middle. A few simple examples, shown in table 16.5, make this clear. Note how the odds equal 1 when the probability of success and failure are equal. The fact that, e.g., the odds are 1/9 vs. 9 for success probabilities of 0.1 and 0.9 respectively demonstrates how 1.0 can be the "center" of the odds range of 0 to infinity.

Here is one way to think about odds. If the odds are 9 or 9/1, which is often written as 9:1 and read 9 to 1, then this tells us that for every nine successes there is one failure on average. For odds of 3:1, for every 3 successes there is one failure on average. For odds equal to 1:1, there is one failure for each success on average. For odds of less than 1, e.g., 0.25, write it as 0.25:1 then multiply the numerator and denominator by whatever number gives whole numbers in the answer. In this case, we could multiple by 4 to get 1:4, which indicates that for every one success there are four failures on average. As a final example, if the odds are 0.4, then this is 0.4:1 or 2:5 when I multiply by 5/5, so on average there will be five failures for every two successes.

To calculate probability, p, when you know the odds use the formula

$$p = \frac{\text{odds}}{1 + \text{odds}}.$$

> **The odds of success is defined as the ratio of the probability of success to the probability of failure. It ranges from 0 to infinity.**

The **log odds** of success is defined as the natural (i.e., base e, not base 10) log of the odds of success. The concept of log odds is very hard for humans to understand, so we often "undo" the log odds to get odds, which are then more interpretable. Because the log is a natural log, we undo log odds by taking Euler's constant (e), which is approximately 2.718, to the power of the log odds. For example, if the log odds are 1.099, then we can find $e^{1.099}$ as exp(1.099) in most computer languages or in Google search to find that the odds are 3.0 (or 3:1). Alternatively, in Windows calculator (scientific view) enter 1.099, then click the Inv (inverse) check box, and click the "ln" (natural log) button. (The "exp" button is *not* an equivalent calculation in Windows calculator.) For your handheld calculator, you should look up how to do this using 1.099 as an example.

The log odds scale runs from $-\infty$ to $+\infty$ with 0.0 in the middle. So zero represents the situation where success and failure are equally likely, positive log odds values represent a greater probability of success than failure, and negative log odds values represent a greater probability of failure than success. Importantly, because log odds of $-\infty$ corresponds to probability of success of 0, and log odds of $+\infty$ corresponds to probability of success of 1, the model "log odds of success equal eta" cannot give invalid probabilities as predictions for any combination of explanatory variables.

It is important to note that in addition to population parameter values for an ideal model, odds and log odds are also used for observed percent success. E.g., if we observe 5/25=20% successes, then we say that the (observed) odds of success is 0.2/0.8=0.25.

> **The log odds of success is simply the natural log of the odds of success. It ranges from minus infinity to plus infinity, and zero indicates that success and failure are equally likely.**

As usual, any model prediction, which is the probability of success in this situation, applies for all subjects with the same levels of all of the explanatory variables. In logistic regression, we are assuming that for any such group of subjects the prob-

ability of success, which we can call p, applies individually and independently to each of the set of similar subjects. These are the conditions that define a binomial distribution (see section 3.9.1). If we have n subjects all with with the same level of the explanatory variables and with predicted success probability p, then our error model is that the outcomes will follow a random binomial distribution written as Binomial(n,p). The mean number of successes will be the product np, and the variance of the number of success will be $np(1 - p)$. Note that this indicates that there is no separate variance parameter (σ^2) in a logistic regression model; instead the variance varies with the mean and is determined by the mean.

The error model for logistic regression is that for each fixed combination of explanatory variables the distribution of success follows the binomial distribution, with success probability, p, determined by the means model.

16.3.2   Example and EDA for logistic regression

The example that we will use for logistic regression is a simulated dataset (LRex.dat) based on a real experiment where the experimental units are posts to an Internet forum and the outcome is whether or not the message received a reply within the first hour of being posted. The outcome variable is called "reply" with 0 as the failure code and 1 as the success code. The posts are all to a single high volume forum and are computer generated. The time of posting is considered unimportant to the designers of the experiment. The explanatory variables are the length of the message (20 to 100 words), whether it is in the passive or active voice (coded as an indicator variable for the "passive" condition), and the gender of the fake first name signed by the computer (coded as a "male" indicator variable).

Plotting the outcome vs. one (or each) explanatory variable is not helpful when there are only two levels of outcome because many data points end up on top of each other. For categorical explanatory variables, cross-tabulating the outcome and explanatory variables is good EDA.

For quantitative explanatory variables, one reasonably good possibility is to break the explanatory variable into several groups (e.g., using Visual Binning in SPSS), and then to plot the mean of the explanatory variable in each bin vs. the

observed fraction of successes in that bin. Figure 16.1 shows a binning of the
length variable vs. the fraction of successes with separate marks of "0" for active
vs. "1" for passive voice. The curves are from a non-parametric smoother (loess)
that helps in identifying the general pattern of any relationship. The main things
you should notice are that active voice messages are more likely to get a quick
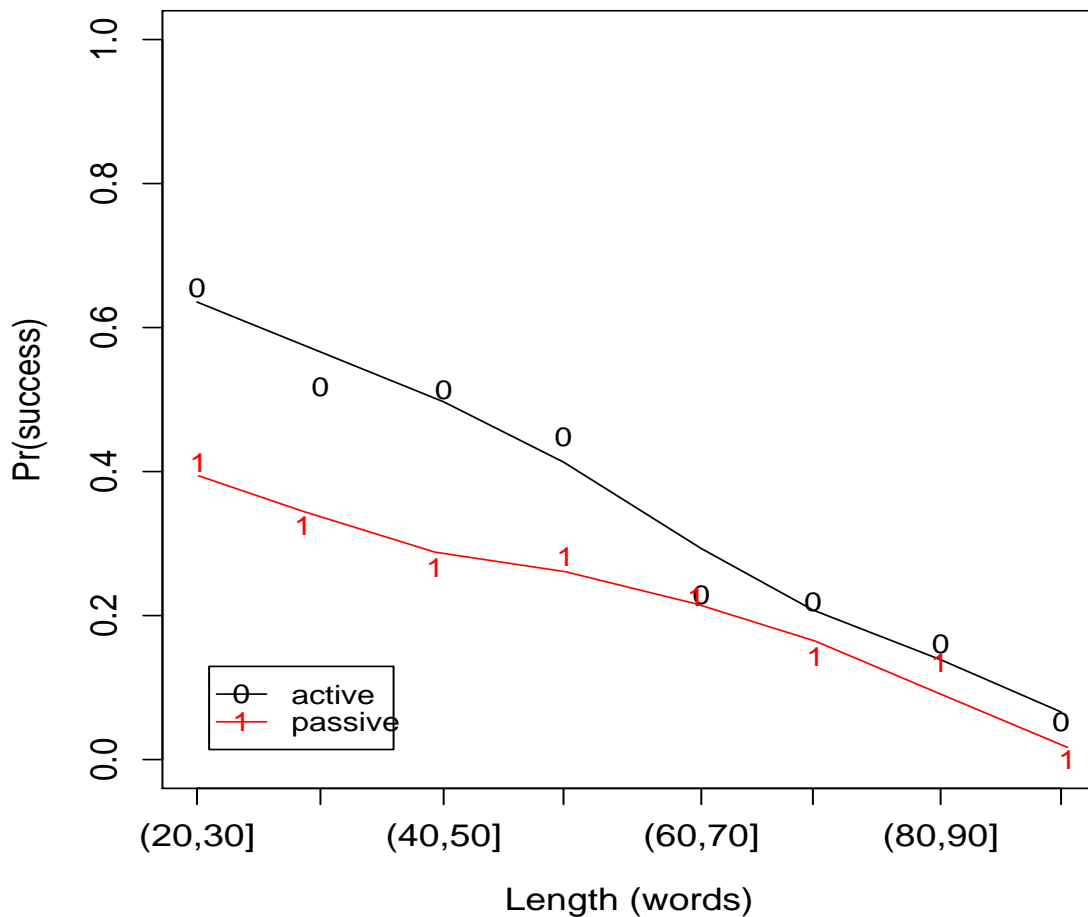reply, as are shorter messages.

Figure 16.1: EDA for forum message example.

EDA for continuous explanatory variables can take the form of categorizing the continuous variable and plotting the fraction of success vs. failure, possibly separately for each level of some other categorical explanatory variable(s).

16.3.3 Fitting a logistic regression model

The means model in logistic regression is that

$$\text{logit}(S) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

For any continuous explanatory variable, x_i, at any fixed levels of all of the other explanatory variables this is linear on the logit scale. What does this correspond to on the more natural probability scale? It represents an "S" shaped curve that either rises or falls (monotonically, without changing direction) as x_i increases. If the curve is rising, as indicated by a positive sign on β_i, then it approaches $\Pr(S){=}1$ as x_i increases and $\Pr(S){=}0$ as x_i decreases. For a negative β_i, the curve starts near $\Pr(S){=}1$ and falls toward $\Pr(S){=}0$. Therefore a logistic regression model is only appropriate if the EDA suggest a monotonically rising or falling curve. The curve need not approach 0 and 1 within the observed range of the explanatory variable, although it will at some extreme values of that variable.

It is worth mentioning here that the magnitude of β_i is related to the steepness of the rise or fall, and the value of the intercept relates to where the curve sits left to right.

The fitting of a logistic regression model involves the computer finding the best estimates of the β values, which are called b or B values as in linear regression. Technically logistic regression is a form of generalized (not general) linear model and is solved by an iterative method rather than the single step (closed form) solutions of linear regression.

In SPSS, there are some model selection choices built-in to the logistic regression module. These are the same as for linear regression and include "Enter" which just includes all of the explanatory variables, "Backward conditional (stepwise)" which starts with the full model, then drops possibly unneeded explanatory variables one at a time to achieve a parsimonious model, and "Forward conditional

Dependent Variable Encoding

Original Value	Internal Value
Not a quick reply	0
Got a quick reply	1

Table 16.6: Dependent Variable Encoding for the forum example.

(stepwise)" which starts with a simple model and adds explanatory variables until nothing "useful" can be added. Neither of the stepwise methods is guaranteed to achieve a "best" model by any fixed criterion, but these model selection techniques are very commonly used and tend to be fairly good in many situations. Another way to perform model selection is to fit all models and pick the one with the lowest AIC or BIC.

The results of an SPSS logistic regression analysis of the forum message experiment using the backward conditional selection method are described here. A table labeled "Case Processing Summary" indicates that 500 messages were tested. The critical "Dependent Variable Encoding" table (Table 16.6) shows that "Got a quick reply" corresponds to the "Internal Value" of "1", so that is what SPSS is currently defining as success, and the logistic regression model is estimating the log odds of getting a quick reply as a function of all of the explanatory variables. *Always check the Dependent Variable Encoding.* You need to be certain which outcome category is the one that SPSS is calling "success", because if it is not the one that you are thinking of as "success", then all of your interpretations will be backward from the truth.

The next table is Categorical Variables Codings. Again *checking this table is critical* because otherwise you might interpret the effect of a particular categorical explanatory variable backward from the truth. The table for our example is table 16.7. The first column identifies each categorical variable; the sections of the table for each variable are interpreted entirely separately. For each variable with, say k levels, the table has k lines, one for each level as indicated in the second column. The third column shows how many experimental units had each level of the variable, which is interesting information but not the critical information of the table. The critical information is the final $k - 1$ columns which explain the coding for each of the $k - 1$ indicator variables created by SPSS for the variable. In our example, we made the coding match the coding we want by using the Categorical button and then selecting "first" as the "Reference category". Each

		Frequency	Parameter coding (1)
Male gender?	Female	254	.000
	Male	246	1.000
Passive voice?	Active voice	238	.000
	Passive voice	262	1.000

Table 16.7: Categorical Variables Codings for the forum example.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.597	8	0.800
2	4.230	8	0.836

Table 16.8: Hosmer-Lemeshow Goodness of Fit Test for the forum example.

of the $k-1$ variables is labeled "(1)" through "(k-1)" and regardless of how we coded the variable elsewhere in SPSS, the level with all zeros is the "reference category" (baseline) for the purposes of logistic regression, and each of the k-1 variables is an indicator for whatever level has the Parameter coding of 1.000 in the Categorical Variables Coding table. So for our example the indicators indicate male and passive voice respectively.

> **Correct interpretation of logistic regression results in SPSS critically depends on correct interpretation of how both the outcome and explanatory variables are coded.**

SPSS logistic regression shows an uninteresting section called "Block 0" which fits a model without any explanatory variables. In backward conditional model selection Block 1 shows the results of interest. The numbered steps represent different models (sets of explanatory variables) which are checked on the way to the "best" model. For our example there are two steps, and therefore step 2 represents the final, best model, which we will focus on.

One result is the **Hosmer and Lemeshow Test of goodness of fit**, shown in Table 16.8. We only look at step 2. The test is a version of a goodness-of-fit chi-square test with a null hypothesis that the data fit the model adequately. Therefore, a p-value *larger* than 0.05 suggests an adequate model fit, while a small p-value indicates some problem with the model such as non-monotonicity, variance inappropriate for the binomial model at each combination of explanatory variables, or the need to transform one of the explanatory variables. (Note that Hosmer and Lemeshow have deprecated this test in favor of another more recent one, that is not yet available in SPSS.) In our case, a p-value of 0.836 suggests no problem with model fit (but the test is not very powerful). In the event of an indication of lack of fit, examining the Contingency Table for Hosmer and Lemeshow Test may help to point to the source of the problem. This test is a substitute for residual analysis, which in raw form is uninformative in logistic regression because there are only two possible values for the residual at each fixed combination of explanatory variables.

> **The Hosmer-Lemeshow test is a reasonable substitute for residual analysis in logistic regression.**

The Variables in the Equation table (Table 16.9) shows the estimates of the parameters, their standard errors, and p-values for the null hypotheses that each parameter equals zero. Interpretation of this table is the subject of the next section.

16.3.4   Tests in a logistic regression model

The main interpretations for a logistic regression model are for the parameters. Because the structural model is

$$\text{logit}(S) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

the interpretations are similar to those of ordinary linear regression, but the linear combination of parameters and explanatory variables gives the log odds of success rather than the expected outcome directly. For human interpretation we usually convert log odds to odds. As shown below, it is best to use the odds scale for interpreting coefficient parameters. For predictions, we can convert to the probability scale for easier interpretation.

	B	S.E.	Wald	df	Sig.	Exp(B)
length	-0.035	0.005	46.384	1	<0.005	0.966
passive(1)	-0.744	0.212	12.300	1	<0.005	0.475
Constant	1.384	0.308	20.077	1	<0.005	3.983

Table 16.9: Variables in the equation for the forum message example.

The coefficient estimate results from the SPSS section labeled "Variables in the Equation" are shown in table 16.9 for the forum message example. It is this table that you should examine to see which explanatory variables are included in the different "steps", i.e., which means model corresponds to which step. Only results for step 2 are shown here; step 1 (not shown) indicates that in a model including all of the explanatory variables the p-value for "male" is non-significant (p=0.268).

This model's prediction equation is

$$\text{logit}(S) = \beta_0 + \beta_{\text{length}}(\text{length}) + \beta_{\text{passive}}(\text{passive})$$

and filling in the estimates we get

$$\widehat{\text{logit}}(S) = 1.384 - 0.035(\text{length}) - 0.744(\text{passive}).$$

The intercept is the average log odds of success when all of the explanatory variables are zero. In this model this is the meaningless extrapolation to an active voice message with zero words. If this were meaningful, we could say that the estimated log odds for such messages is 1.384. To get to a more human scale we take $\exp(1.384)=e^{1.384}$ which is given in the last column of the table as 3.983 or 3.983:1. We can express this as approximately four successes for every one failure. We can also convert to the probability scale using the formula $p = \frac{3.983}{1+3.983} = 0.799$, i.e., an 80% chance of success. As usual for an intercept, the interpretation of the estimate is meaningful if setting all explanatory variables to zero is meaningful and is not a gross extrapolation. Note that a zero log odds corresponds to odds of $e^0 = 1$ which corresponds to a probability of $\frac{1}{1+1} = 0.5$. Therefore it is almost never valid to interpret the p-value for the intercept (constant) in logistic regression because it tests whether the probability of success is 0.5 when all explanatory variables equal zero.

> The intercept estimate in logistic regression is an estimate of the log
> odds of success when all explanatory variables equal zero. If "all
> explanatory variables are equal to zero" is meaningful for the problem,
> you may want to convert the log odds to odds or to probability. You
> should ignore the p-value for the intercept.

For a k-level categorical explanatory variable like "passive", SPSS creates $k-1$ indicator variables and estimates $k-1$ coefficient parameters labeled $B_{\text{x}(1)}$ through $B_{\text{x(k-1)}}$. In this case we only have $B_{\text{passive}(1)}$ because $k = 2$ for the passive variable. As usual, $B_{\text{passive}(1)}$ represents the effect of increasing the explanatory variable by one-unit, and for an indicator variable this is a change from baseline to the specified non-baseline condition. The only difference from ordinary linear regression is that the "effect" is a change in the log odd of success.

For our forum message example, the estimate of -0.744 indicates that at any fixed message length, a passive message has a log odds of success 0.744 lower than a corresponding active message. For example, if the log odds of success for active messages for some particular message length is 1.744, then the log odds of success for passive messages of the same length is 1.000.

Because log odds is hard to understand we often rewrite the prediction equation as something like

$$\widehat{\text{logit}}(S) = B_{0L} - 0.744(\text{passive})$$

where $B_{0L} = 1.384 - 0.035L$ for some fixed message length, L. Then we exponentiate both sides to get

$$\widehat{\text{odds}}(S) = e^{B_{0L}}e^{-0.744(\text{passive})}.$$

The left hand side of this equation is the estimate of the odds of success. Because $e^{-0.744} = 0.475$ and $e^0 = 1$, this says that for active voice $\widehat{\text{odds}}(S) = e^{B_{0L}}$ and for passive voice $\widehat{\text{odds}}(S) = 0.475e^{B_{0L}}$. In other words, at any message length, compared to active voice, the odds of success are *multiplied* (not added) by 0.475 to get the odds for passive voice.

So the usual way to interpret the effect of a categorical variable on a binary outcome is to look at "exp(B)" and take that as the multiplicative change in odds when comparing the specified level of the indicator variable to the baseline level.

If B=0 and therefore exp(B) is 1.0, then there is no effect of that variable on the outcome (and the p-value will be non-significant). If exp(B) is greater than 1, then the odds increase for the specified level compared to the baseline. If exp(B) is less than 1, then the odds decrease for the specified level compared to the baseline. In our example, 0.475 is less than 1, so passive voice, compared to active voice, lowers the odds (and therefore probability) of success at each message length.

It is worth noting that multiplying the odds by a fixed number has very different effects on the probability scale for different baseline odds values. This is just what we want so that we can keep the probabilities between 0 and 1. If we incorrectly claim that for each one-unit increase in x probability rises, e.g., by 0.1, then this becomes meaningless for a baseline probability of 0.95. But if we say that, e.g., the odds double for each one unit increase in x, then if the baseline odds are 0.5 or 2 or 9 (with probabilities 0.333, 0.667 and 0.9 respectively) then a one-unit increase in x changes the odds to 1, 4 and 18 respectively (with probabilities 0.5, 0.8, and 0.95 respectively). Note that all new probabilities are valid, and that a doubling of odds corresponds to a larger probability change for midrange probabilities than for more extreme probabilities. This discussion also explains why you cannot express the interpretation of a logistic regression coefficient on the probability scale.

> **The estimate of the coefficient for an indicator variable of a categorical explanatory variable in a logistic regression is in terms of exp(B). This is the *multiplicative* change in the odds of success for the named vs. the baseline condition when all other explanatory variables are held constant.**

For a quantitative explanatory variable, the interpretation of the coefficient estimate is quite similar to the case of a categorical explanatory variable. The differences are that there is no baseline, and that x can take on any value, not just 0 and 1. In general, we can say that the coefficient for a given continuous explanatory variable represents the (additive) change in log odds of success when the explanatory variable increases by one unit with all other explanatory variables held constant. It is easier for people to understand if we change to the odds scale. Then exp(B) represents the *multiplicative* change in the odds of success for a one-unit increase in x with all other explanatory variables held constant.

For our forum message example, our estimate is that when the voice is fixed at either active or passive, the log odds of success (getting a reply within one

hour) decreases by 0.035 for each additional word or by 0.35 for each additional ten words. It is better to use exp(B) and say that the odds are multiplied by 0.966 (making them slightly smaller) for each additional word.

It is even more meaningful to describe the effect of a 10 word increase in message length on the odds of success. Be careful: you can't multiply exp(B) by ten. There are two correct ways to figure this out. First you can calculate $e^{-0.35} = 0.71$, and conclude that the odds are multiplied by 0.71 for each additional ten words. Or you can realize that if for each additional word, the odds are multiplied by 0.966, then adding a word ten times results in multiplying the odds by 0.966 ten times. So the result is $0.966^{10} = 0.71$, giving the same conclusion.

The p-value for each coefficient is a test of $\beta_x = 0$, and if $\beta_x = 0$, then when x goes up by 1, the log odds go up by 0 and the odds get multiplied by exp(0)=1. In other words, if the coefficient is not significantly different from zero, then changes in that explanatory variable do not affect the outcome.

For a continuous explanatory variable in logistic regression, exp(B) is the multiplicative change in odds of success for a one-unit increase in the explanatory variable.

16.3.5   Predictions in a logistic regression model

Predictions in logistic regression are analogous to ordinary linear regression. First create a prediction equation using the intercept (constant) and one coefficient for each explanatory variable (including $k - 1$ indicators for a k-level categorical variable). Plug in the estimates of the coefficients and a set of values for the explanatory variables to get what we called η, above. This is your prediction of the log odds of success. Take $\exp(\eta)$ to get the odds of success, then compute $\frac{\text{odds}}{1+\text{odds}}$ to get the probability of success. Graphs of the probability of success vs. levels of a quantitative explanatory variable, with all other explanatory variable fixed at some values, will be S-shaped (or its mirror image), and are a good way to communicate what the means model represents.

For our forum messages example, we can compute the predicted log odds of success for a 30 word message in passive voice as $\eta = 1.384 - 0.035(30) - 0.744(1) =$

−0.41. Then the odds of success for such a message is exp(-0.41)=0.664, and the probability of success is 0.664/1.664=0.40 or 40%.

Computing this probability for all message lengths from 20 to 100 words separately for both voices gives figure 16.2 which is a nice summary of the means model.
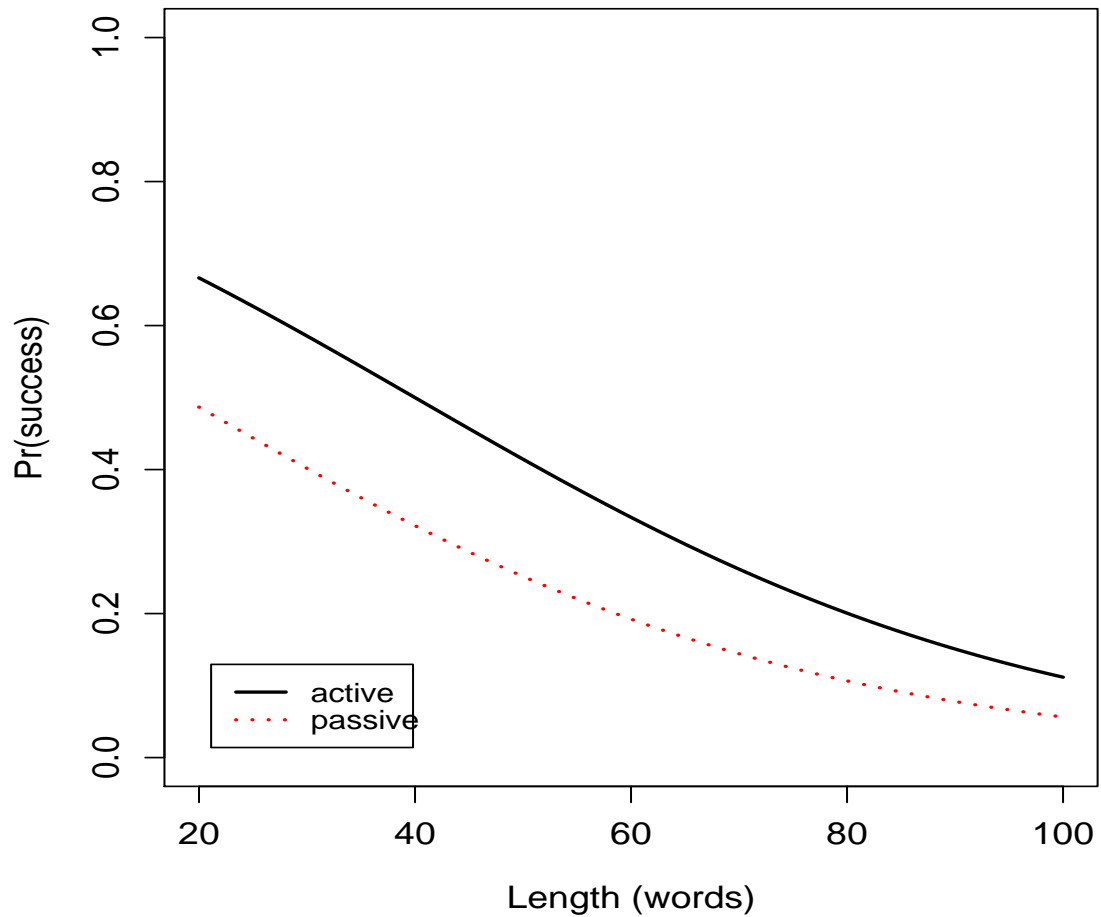

Figure 16.2: Model predictions for forum message example.

> **Prediction of probabilities for a set of explanatory variables involves calculating log odds from the linear combination of coefficient estimates and explanatory variables, then converting to odds and finally probability.**

16.3.6   Do it in SPSS

In SPSS, Binary Logistic is a choice under Regression on the Analysis menu. The dialog box for logistic regression is shown in figure 16.3. Enter the dependent variable. In the "Covariates" box enter both quantitative and categorical explanatory variables. You do *not* need to manually convert k-level categorical variables to indicators. Select the model selection method. The default is to "Enter" all variables, but you might want to switch to one of the available stepwise methods. You should always select "Hosmer-Lemeshow goodness-of-fit" under Options.
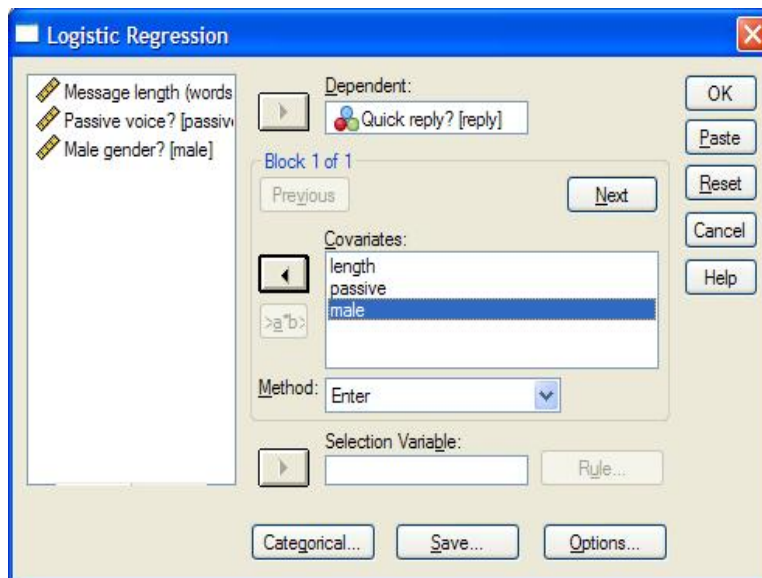

Figure 16.3: SPSS dialog box for logistic regression.

If you have any categorical explanatory variables listed in the "Covariates" box, click on "Categorical" to open the dialog box shown in figure 16.4. Move only the

categorical variables over to the "Categorical Covariates" box. The default is for SPSS to make the last category the baseline (reference) category. For variables that are already appropriately named indicator variables, like passive and male in our example, you will want to change the "Reference Category" to "First" to improve the interpretability of the coefficient tables. Be sure to click the "Change" button to register the change in reference category.
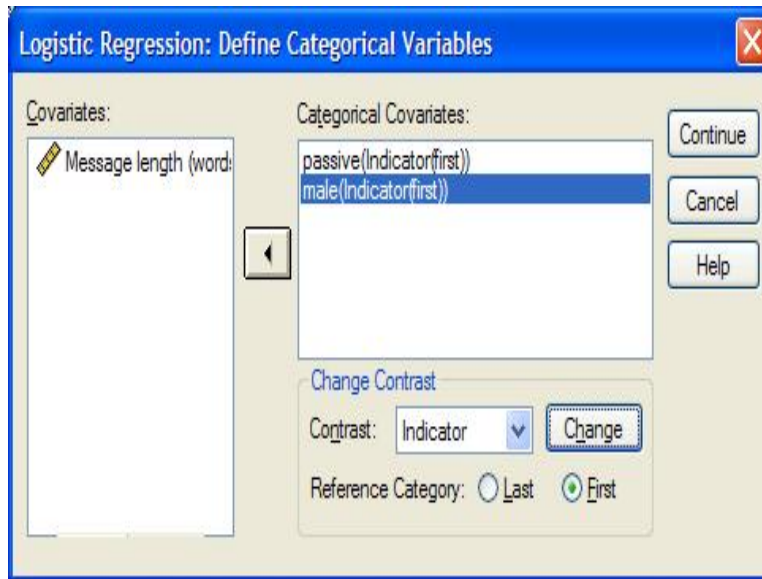

Figure 16.4: SPSS Categorical Definition dialog box for logistic regression.

The interpretation of the SPSS output is shown in the preceding sections.