# This is Professor Rogers.

In this presentation we look at the use of categorical dependent variables using logistic models.

The example problem comes from the National Survey of Drug Use and Health. We will look at a variable that identifies whether a person has a severe mental illness. It is a dichotomous response—0 for no and 1 for yes. The independent variables, also dichotomous, measure whether a person lives at or below the poverty level and whether the person lives in a non-metropolitan area.

# The old solution to this problem was to use ordinary least squares regression.

In this model, the predicted value for each observation becomes the probability that the person has a severe mental illness. Here we see a model that finds that poverty is statistically significant and non-metro residence is not. The model explains only 0.4% of the variance.

Probabilities range between 0 and 1—0 if the event did not occur and 1 if the event happens every time. The problem with the regression approach is that predicted values can be less than 0 and greater than 1, which makes no sense. In this example, you see that the minimum value is less than 0.

# Enter logistic regression.

Logistic regression transforms the distribution of predicted probabilities into an S-shaped curve, so named because the bends in the curve are similar to the way the letter S bends.

In this example, we try to predict whether a person will pass an exam by the hours studying. See the dots on the top and bottom of the y-axis. Those are the actual values of the result—1 if the person passes and 0 if the person fails. When we calculate the probability of passing based on the hours spent studying, we get the S-shaped curve of probabilities seen here.

Bounded by zero on the bottom and one on the top, this distribution of expected values conforms to our expectations of what the data really is like.

The dependent variable in logistic regression is a transformation of the dichotomous variable into a log-odds ratio.

We denote the dependent variable as P(t), which is the probability of event t. We call this the logit or log odds ratio.

The odds ratio part is the probability of an event, here denoted by the letter p, divided by 1 minus the probability.

If something happened 20% of the time, then the odds ratio associated with the event is .25. .25, as you know, is 1 over 4. For every time the event happens, it does not happen 4 times.

We then take the natural log of the odds ratio, which in this case gives us a log odds-ratio of approximately -1.39.

# Here’s the full equation for logistic regression shown two different ways.

Note what happens in the second equation when we unwind the log odds ratio to move back to the odds ratio. See that small letter t in the denominator—that’s where our equation is. It’s not a very intuitive number, but we can learn how to interpret it fairly quickly.

In SPSS, the type of logistic regression with a 0/1 response variable is known as binary logistic regression. The path for the logistic regression command is Analyze > Regression > Binary Logistic.

The same path is present in PSPP, but the regression options are not as many as SPSS.

In the SPSS command window, the dependent variable goes on the top right side, where we are used to seeing it, and the independent variables are entered beneath it in the box labelled Covariates.

In PSPP, the box uses dependent variable and the more conventional term independent variables.

# There’s a number of pieces to the output.

We want to go to Block 1, or the last block, which is the final result. There are four tables that are part of this block.

The bottom box titled “Variables in the Equation” contains the parameters for fitting the curve.

The first column contains the betas for the equation. These estimates give us direction and unstandardized magnitudes. The unstandardized magnitudes are hard to interpret given that we are fitting a curve, not a line, so it is often not customary to do so.

The final column, labeled Exp(B), gives us a relative interpretation of their strength. The Exp(B) is called the odds ratio, or the exponentiated beta, and refers to the likelihood of an event relative to the reference value. Here’s the rule for interpretating odds ratios.

Unlike other parameters, where the null hypothesis is 0, the null hypothesis of the odds ratio is 1.00, which means there is no relationship. Think of this like horse racing: If you placed a bet on a race in which the odds were 1 to 1, you would get back one dollar for every dollar you wagered. There’s really no point to that.

The odds ratio is interpreted relative to a reference category. A positive relationship is signaled by a number above one. The odds ratio for whether a household is in poverty is 1.72. This means that the event is 1.72X more likely to occur than the reference value, which is whether the household is above the poverty line. We can also say it is 72% more likely, which is determined by subtracting the null hypothesis of 1 from the odds ratio, then multiplying by one hundred.

A negative relationship has an odds ratio of less than 1. The odds ratio for non-metropolitan locations is .99. I would way that the event was 1% less likely to occur than the reference category, which is metropolitan locations. I get 1% by subtracting .99 from the null hypothesis of 1.00, then multiplying by 100.

# Three additional comments about the rules for interpretation.

First, odds ratios give us relative values. A few years ago I published a paper in which I reported that young boys were 76% more likely to need habilitative therapy than girls. This gives us an image of large numbers of preschool boys running around with speech defects. However, among all children, only 3.4% have need of therapy. The 76% gap must be understood in terms of a very small base.

Second, logistic regression can be used with a block of dummy variables to represent a variable with three or more variables. We hold out one category, whose value goes to the intercept, and then include the other values and treat them as relativistic values relative to the reference category. For example, with the variable Region, I might hold back the region West and then add dummy variables for Northeast , Midwest, and South. The odds ratios are interpreted as comparison relative to the reference category.

Finally, we can use either categorical or numeric variables as independent variables. However, we need to be extra cautious with numeric variables. With numeric variables, the odds ratio means an increase for each point increase of the independent variable. If we are using a variable with large values, we might get ratios whose difference from the null hypothesis is tiny. In that case, we might consider creating a categorical variable out of our numeric variable to highlight our results better.

The column Sig. is the statistical significance value. We read it the same as we have other significance values: a p-value of less than .05 indicates statistical significance using a 95% confidence interval. The three previous columns are used to calculate the p-value. Instead of t, which was used in OLS regression, the logistic model uses the Wald chi-square, a slight departure from the chi-square distribution that we dealt with in contingency tables.

The model summary table gives us goodness-of-fit statistics.

At the bottom we see something that we have not seen before—iterations. Iterations have to do with the complexity of the calculation. Calculus is required to solve the problem, and computers can’t do calculus. Instead, the computer makes an estimate, and then tweaks it. It keeps tweaking the estimate until the error is so minimal that we don’t care, and then it stops. Each time around it is called an iteration. At each iteration the computer calculates a log likelihood function. Some computer packages give us information about each iteration—others give us only the final iteration.

The log likelihood is loosely analogous to the sum of squares. When I tried to find an easy interpretation of the number, I kept finding stats books that said things like you don’t need to know—it is just a number for comparative purposes. If we dissect the formula, though, we can get a general idea of what it is about.

Here on the left side of the equation, we have the function for P(t), which is a probability. It is kind of like our predicted value.

On the right side of the equation, we have one minus the probability. You remember that the odds ratio was built on the probability divided by one minus the probability, then logged. What we essentially are doing is finding all the different probabilities of the event occurring or not occurring, logging them, and then expressing them in terms of y.

The log likelihood by itself it means nothing, but it is quite useful when comparing similar equations. Iteration 0 or what SPSS calls Block 0 is our baseline—the model before we enter any variables. It is like the total sum of squares in OLS regression. The ratio of the -2 log likelihood of the final iteration to the -2 log likelihood of the first iteration is the likelihood chi-square of the model. This chi-square is similar in nature and use to the Pearson’s chi-square we have with contingency tables. In this example, the likelihood chi-square for the model is 138.887, which with two degrees of freedom has a p-value of .000. This tells us the model as a whole is statistically significant.

Log likelihood statistics and chi-squares do not give us an intuitive interpretation of the model’s fit, so there is a need for something like an R-square. Logistic models respond to this need with something called the pseudo-R2.

SPSS reports two different pseudo R-square statistics built around efforts to compare the log-likelihood ratios of the initial and final iterations. The Cox & Snell number cannot reach the value of one. The Nagelkerke number will usually generate a larger pseudo-R-square Most statisticians will say that neither of these pseudo-R-squares can be regarded as equivalent to the R-square of OLS regression models, but they can be used to compare the results of one model to another to determine which has the best fit in a relative sense.

# Another way to discuss the fit of a logistic model is to use a classification table,

which contrasts the actual value (0 or 1) with the predicted value (0 or 1). The predicted model is based on the probabilities determined by the logistic model: An event is regarded as predictive and scored as a 1 if its probability is .5 or greater and not predicted and scored as a 0 if it is less than .5.

Let’s look at this table closely. In our model, the percent correctly predicted is 93.3%.

However, when we look at the guts of the table. We have a problem. Specificity is the number of non-events correctly predicted. In this model, there were 36,497 non-events, and we picked every one of them right to have a specificity rate of 100%.

Sensitivity is the number of events correctly predicted. Here’s the problem: Of 2,636 events, none were correctly picked.

What the model did was simply score every event as if it did not happen. Why did it do this? It’s because the event occurs only 6.7% of the time—the probability of the event occurring is so low that the computer simply decided that the best fit would be to predict non-occurrence of the event in every instance.

We can change this by changing our assumption about the classification table. In the options box, the cutoff point for the classification table can be found near the lower right hand corner of the box. By default, it is set to .5. We can change this to any number between 0 and 1. Since the event occurs 6.7% of the time, we can change the cutoff point to .067 and rerun the model.

Here is the result. The equation of the model does not change, but the classification table does. Our percentage correct drops to 76.6%, but you see inside the table, the sensitivity score improves to 30.1%.

The question is when should we change the cutoff point and when should we not. Drug trials might prefer to keep the .5 threshold. The reason has to do with the severity of a non-event. Let’s suppose I am testing a new drug to cure cancer—1 means the patient survives, and 0 means the patient dies. A drug that works only 6.7% of the time isn’t very effective—people may be because of my new drug. If you view your event in this way, you might want the first classification table.

Human behavior, however, is not so easily described, and some events are rare. The social sciences, including criminal justice and public health, are more likely to change the cutoff point to reflect the actual likelihood of the event. If you view your event in this way, you would want the second classification table.

Statisticians sometimes apply additional techniques to maximize sensitivity and specificity, the most important of these being the use of a ROC curve, but we will not deal with that issue in this presentation.

# Let’s summarize how the characteristics of an inferential statistic

are present in a binary logistic regression.

First, the sign of the beta is the direction.

The beta provides an unstandardized magnitude but it is hard to interpret. The odds ratio is often used instead, but it is a relative magnitude, not a standardized one. There is no standardized magnitude, so terms like weak, moderate, and strong cannot be applied.

Statistical significance is present.

Model performance is complicated. The pseudo R-square is relativistic—it can compare the performance of one equation to another but it is not standardized like the R-square in OLS regression. The classification table gives a very precise measure of accuracy but is unstandardized. In fact, the cut points can even be changed.

# The model that we have looked at so far deals with a response that is binary, 0 and 1.

There are some additional models that can be used when we have more than two categories in our dependent variable. In this example, I will go to the National Survey of Drug Use and Health and look at the entire classification of people by mental illness, ranging from no mental illness in the past year to serious mental illness. The response that we just looked at was simply the highest level of mental illness in this variable—now let’s look at all four responses at the same time. We will look at two different techniques—multinomial logistic regression and ordinal logistic regression. These techniques can only be done in SPSS.

Multinomial regression can be used for either nominal or ordinal data. Its path is Analyze > Regression > Multinomial Logistic.

Here you see the parameter estimates.

It makes one response the reference category. Here it chose serious mental illness.

It gives us the results of three different binary logistic regressions side by side. In each we are comparing each of the categories to whether the person had a serious mental illness. For example, equation 1 is a binary logistic regression comparing whether a person did not have a mental illness, which gets a value of 1, to whether the person had a serious mental illness, which gets a value of 0.

The output is hard to read. The best way to understand what is going on is to lay the three equations side by side and then compare coefficients. Here we see that non-metro residence is not significant in any of the three comparisons. Poverty always has a negative relationship, meaning that it is less likely to be present when a person does not have a serious mental illness. However, in this model you not that the negative relationship is strongest when the comparison involves not having a mental illness in the past year.

The output gives us model-performance statistics across all three categorical comparisons combined. The output is all statistics with which we are already familiar.

If our variable is ordinal, we can take advantage of the ordering of the values using what is called Ordinal Logistic Regression. It’s path is Analyze > Regression > Ordinal.

The output gives us a single equation describing how the likelihood of the event changes as we increase by category.

Here’s the parameters for the independent variables. Note that we do not get an odds ratio with this output in ordinal logistic regression and there is no option to give us a standardized or relativized number in our software package—we have to use the parameter estimates. I should note that some other statistical software packages have a way to provide what is sometimes called the “relative risk ratio,” which acts like an odds ratio.

Instead of an intercept or constant, we get thresholds or cut points. What happens with ordinal logistic regression is that we have different intercepts representing where we are in the categories. Each threshold is a constant that falls between two categories. For example, the threshold of 2.091 is the constant that separates category 0 from category 1.

The goodness-of-fit and pseudo-R-square statistics are already familiar to us. There is no classification table in our software package.

Most of this presentation dealt with binary logistic regression. I also made comments about two other kinds of logistic regressions—multinomial and ordinal—which deal respectively with nominal and ordinal dependent variables.