Week 15 Class Notes
- Go over test questions
- Introduction to Logistic Regression
- Final Paper requirements
- Affinity group work

I.   Logistic Regression

A. Definition and Types

1. *Binomial (or binary) logistic regression* is a form of multivariate regression which is used *when the dependent variable is a dichotomous categorical variable and the independent variables are of any type*. The DV (e.g. cured/not cured; happy/not happy; used service/didn't use service) is coded as a dummy variable (1 = Yes, 0 = No).

2. Although the outcome or dependent variable is dichotomous, logistic regression is similar to multivariate linear regression in that the purpose is to analyze the effects of the predictor variable(s) on the dependent variable, controlling for all other IVs. In this case, it's *the effect of the predictor(s) on whether the outcome occurs or not*.

3. Extensions of the logistic model include *multinomial logistic regression*, used when there are more than two nominal categories in the DV (such as neglected, sexually abused, or physically abused). *Ordinal logistic regression* is used when the DV is ordinal (such as Very happy, moderately happy, moderately unhappy, and very unhappy). In this class we will only discuss the binomial logistic regression model.

B. Typical research questions for binomial logistic regression:

1. "What is the likelihood of having a suicide attempt, controlling for number of previous suicide gestures, severity of depression, ethnicity, gender?"

2. "Does the successful completion of a substance abuse treatment program depend on higher levels of motivation? What role do ethnicity, gender and type of drug play?"

3. "Do support groups for released felons prevent being re-arrested, controlling for number of previous arrests, seriousness of previous crime, and ethnicity?"

4. "Can reunification home after a child abuse report be predicted by type of abuse allegation, satisfaction with services, ethnicity, and previous child abuse reports?"

C. The logic of logistic regression

1. Since the DV is not continuous, the assumption of linearity (i.e. that the IV is linearly related to the DV) does not apply as it does for bivariate and multivariate linear regression.

2. Logistic regression estimates *the odds of a certain event occurring* (the DV event), or how likely it is (the odds) that the observed values of the DV may be predicted from the observed values of the IVs. (See below for definitions of odds and odds ratios.)

3. In research articles you might encounter the term "*maximum likelihood.*" This is the statistical procedure used in logistic regression (it doesn't use least squares regression). Maximum likelihood picks the values of the model parameters (e.g. the beta coefficients) that make the observed data "more likely" than they would be with other parameters. The statistical procedure tries several "iterations" until reaching the best fitting parameter estimates. Making interpretations of the basic statistical output does *not* require knowledge of how maximum likelihood works.

Example of Logistic Regression: "The Effects of Employment, Education, Rehabilitation and Seriousness of Offense on Re-Arrest"
You are working as a social worker in a criminal justice probation agency, and wonder what some of the factors are leading to re-arrest of those managed by your agency over the past five years who were convicted and then released. You have data on 1,000 clients from your facility with the following variables:
- Whether or not the client was adjudicated for a second criminal offense
- Seriousness of first offense (felony vs. misdemeanor)
- High school graduate vs. not
- Whether or not client completed a rehabilitation program after the first offense, and
- Employment status after first offense

1. Identify the independent variables and level of measurement.
- Seriousness of first offense (felony vs. misdemeanor)—categorical, nominal (coded as a dummy variable: 0 = misdemeanor; 1 = felony)
- High school graduate vs. not -- categorical, nominal (coded as a dummy variable: 0 = not graduated; 1 = graduated)
- Whether or not client completed a rehabilitation program after the first offense— categorical, nominal (coded as a dummy variable: 0 = no rehab completed; 1 = rehab completed)
- Employment status after first offense—categorical, nominal (coded as a dummy variable: 0 = not employed; 1 = employed)

Note: independent variables in logistic regression can also be continuous. This scenario doesn't happen to have any of them.

2. Identify the dependent variable and level of measurement.
- Re-arrested vs. not re-arrested – categorical, nominal (coded as a dummy variable: 0 = not re-arrested; 1 = re-arrested)

3. What is the null hypothesis for the overall model fit?

The overall model does not predict re-arrest. OR, the IVs as a group are not related to being re-arrested.

For the Independent variables?

The separate IVs are not related to the likelihood of re-arrest.

4. What is the alternative hypothesis for the overall model fit?

The overall model predicts the likelihood of re-arrest.

For the independent variables?

Having committed a felony (vs. a misdemeanor), not completing high school, not completing a rehab program, and being unemployed are related to the likelihood of being re-arrested.

5. Why is logistic regression suitable for these data? Identify the alpha level.

Categorical (dichotomous) DV with any type of IVs. Alpha = .05

6. Review SPSS output.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	41.155	4	.000
	Block	41.155	4	.000
	Model	41.155	4	.000

This table shows the results of a Chi-Square test, testing the null that the Model, or the group of IVs taken together, does not predict the likelihood of being re-arrested. (The expected and observed values of the Chi Square analysis are not shown.) The p value is the probability of obtaining this chi-square statistic (41.55) if there is in fact no effect of the independent variables, taken together, on the likelihood or odds of re-arrest. (We're concerned about row three— "Model". Ignore the other rows.) As with other statistics, we compare this to our alpha (.05), and we can reject the null.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	felony	.283	.142	3.997	1	.046	1.327
	highschool	.023	.138	.028	1	.867	1.023
	rehab	-.679	.142	22.725	1	.000	.507
	employ	-.513	.142	13.031	1	.000	.599
	Constant	1.035	.154	45.381	1	.000	2.816

This table shows the independent variable coefficients and related statistics. This requires some explanation.

B is the regression coefficient in logistic regression, but it is not interpreted the same as in linear regression. While still showing the impact of a one unit change in the IV on the DV, controlling for all other variables, this B is a "log odds," e.g. committing a felony increases the log odds of re-arrest by .283. Not a very helpful description!

The last column, **Exp(B)** is the "exponentiated B" (taking the B value and calculating the inverse natural log[1] of B) which is, conveniently, an odds ratio. **Odds** means the probability of an event occurring, divided by the probability of the event not occurring. Example: let's say the probability of being randomly selected in class to answer a question is 1 out of 25, or .04. The probability of <u>not</u> being selected is (1 - .04) = .96 (the same as 24/25). The odds of being selected are .04/.96 = .04167, rounded to .04.

Now let's say you are also in another class with 20 students. Your odds of being randomly selected there are 1/20 divided by 19/20, = .05/.95 = .0526, rounded to .05. An **Odds Ratio** is the odds of an event occurring in one group, divided by the odds of the event occurring in another group. So your odds ratio of being selected in the first class vs. the other is .04/.05 = 0.8, which means the odds of being selected in the first class are 80% that of the second class. If you reversed the numerator and denominator to .05/.04 = 1.25, you would have a 25% greater odds of being selected in the second class.

Now let's look at the odds ratio of being re-arrested vs. not re-arrested, as affected by being in the felony group. The odds ratio compares the two groups (re-arrested = 1 in the numerator, and re-arrested = 0 in the denominator) for the felony group, compared to the baseline misdemeanor group. Exp(B) for "felony" is the "one unit change[2]" of felony on the odds of being re-arrested divided by the odds of not being arrested, or 1.327. We can say that having committed a felony

---

[1] The "natural logarithm" of a number x is the power to which e would have to be raised to equal x. For example, the natural log of e itself is 1, since $e^1 = 1$. The natural log of 1 would be 0, since $e^0 = 1$.

[2] If the IV were a continuous variable such as age, the interpretation would be the same—Exp(B) would be the effect on the odds ratio of "re-arrested" by a one-unit change in the continuous variable (or a one-year increase in age)

vs. misdemeanor *increases the odds of re-arrest by 33%*. For "rehab" we can say that having completed rehab *reduces the likelihood (or odds) of being re-arrested by almost 51%.*

An Exp(B) over 1.0 signifies that the independent variable *increases the odds* of the dependent variable occurring (whatever was coded 1 in the DV). An Exp(B) under 1.0 signifies that the independent variable *decreases the odds* of the dependent variable occurring. Another clue about the direction of effect can be seen in B — a negative B coefficient will result in an Exp(B) less than 1.0, and a positive B coefficient will result in an Exp(B) greater than 1.0.

The statistical significance of each B is tested by the **Wald Chi-Square**—testing the null that the B coefficient = 0 (the alternate hypothesis is that it does not = 0). *p* values lower than alpha are significant, leading to rejection of the null.

7. Results—decision to accept/reject null

The Omnibus Test of Model Coefficients shows that the overall model is predictive of re-arrest, X^2 (4) = 41.15, *p* < .001, and the null can be rejected. We can also reject the null that the B coefficients for having committed a felony, completing a rehab program, and being employed are equal to zero—they are statistically significant and predictive of re-arrest. Education level, however, was not found to be predictive of re-arrest. Controlling for other variables, having committed a felony for the first offense increases the odds of being re-arrested by 33% (*p* = .046), compared to having committed a misdemeanor. Completing a rehab program and being employed after the first offense decreases the odds or re-arrest, each by more than 50% (*p* < .001).

8. Discussion of Results

The findings support the use of employment services and rehabilitative programs for those convicted of a first offense. Our study found that those individuals who completed a rehab program (compared to those who did not) and were employed after the first offense (compared to those who were not) were more likely to avoid a second offense during the time period of the study. Since those who committed a felony were more at risk for re-arrest, those programs should be made available to the more serious offenders. One serious limitation of the study is that the sample included all clients seen in the probation agency in the past five years. Clients were not differentiated by when, during those five years, they had committed the first offense. Clients whose first offense occurred later in the study period may not have had sufficient time to allow the effects of educational attainment, employment and rehabilitation programs to affect their re-arrest, and information was not available about clients whose re-arrest may have occurred after the study period. Future research should control for time since first offense. There may also be other variables not included that may be predictive of future offenses, such as ethnicity, community of residence, and substance abuse. Future studies can use more rigorous research designs in order to compare matched groups based on key variables (such as rehab program participation, or type of first offense. For example, is a rehab program equally as effective with both felony and misdemeanor first offenses?). Despite the limitations, this study points to the importance of post-release programs and employment to minimize the risks leading to recidivism.