Stat 324 – Day 33
Inference for Logistic Regression

Last Time: Residuals
- Two types of residuals with “grouped data” (binomial): Pearson residuals and Deviance residuals
  - Help identify unusual constellations (|standardized residual| > 2)
  - Measures of influence include Delta Beta (Minitab) and Cook’s D
- The sum of the squared residuals (akin to SSE) give us measure of lack-of-fit
  - Large test statistic values are evidence that the logistic model is not appropriate
  - Need replication for Pearson and Deviance tests, otherwise use Hosmer-Lemeshow

Correct Classification Rates
One way to assess the performance of the model is see what percentage of observations are correctly classified as “success” and “failure.” One approach is to code a success for observations with estimated probability of success > .5 (or start with overall \( \hat{\pi} \) and try different cut-off values to minimize misclassification errors) and compare this to the actual outcome, reporting the percentage of correct classifications (S,S and F,F over total). Can create a “confusion matrix” with ungrouped data.
- Some suggest comparing this proportion to max\( n_1/n, n_2/n \).
- Also consider cross-validation.

Example 1 Reconsider the Death Penalty study
(a) For the Death Penalty study, store and examine the predicted probabilities for the model using aggravation (Day 31 technology instructions, in JMP edit the formula). How many death penalty verdicts were given out when the constellation had a predicted probability less than .5? How many times was a death penalty verdict not given out when the constellation had a predicted probability greater than .5? So what is the total misclassification rate? (Also record the Deviance statistic and AICc.)

(b) Find the misclassification rate for the model using both aggravation and race of victim. How does this compare? (Also record the Deviance statistic and AICc.)

Statistical Inference
Instead of SSE, we will calculate the sum of the square deviance residuals as a measure of model fit. The “partial F-test” will replaced by a “drop in deviance” test. The reference distribution for a drop in deviance test is a Chi-square distribution with df = difference in number of parameters between the two models. These tests are related to more general class of “likelihood ratio tests” that you will learn about in Stat 426.
- **The overall model utility test** compares model deviance to the “null model” deviance (null model) = predict the same overall probability for all observations in the data set, like comparing to the model with \( \hat{y}_i = \bar{y} \).
- **Significance of individual variables** is tested by comparing the deviance of the model with that variable to the deviance of the model without that variable.
Note: These tests are alternatives to “Wald Tests” which have the more familiar form of estimate/SE, (which when squared follows a Chi-square distribution with df = 1) but are less reliable with smaller sample sizes. See More details Technology instructions.

(c) For the model with aggravation and victim race, find the overall model test/whole model test. What are the hypotheses, test statistic, degrees of freedom, and p-value? What do you conclude from this test?

(d) Is the victim’s race a significant predictor of receiving the death penalty after adjusting for aggravation level of the crime? What is the test statistic, degrees of freedom, and p-value? Is this p-value one-sided or two-sided? What do you conclude from this test?

Example 2: A study determined risk factors for kyphosis, which is severe forward flexion of the spine following corrective spinal surgery. The data are in kyphosis.xls.

(a) Are the data grouped or ungrouped? Do we have much replication?

(b) Fit a logistic regression model using age (in months, at time of operation) as a predictor of whether kyphosis is present. Examine the Lack of Fit tests, what do you conclude? Do any observations stand out as being unusual?

(c) Fit a model that also includes age^2. Compare the Lack of Fit tests – have they improved? Does the quadratic term appear to be significant? What relationship does this model suggest between the probability of kyphosis and age? Does that make sense in this context? Is it consistent with graphs of age for the two groups?

Practice Problem – due 9am Monday
Example 3: Field goal kicking data for the American Football League (AFL) and the National Football League (NFL) for the 1969 season are given in Fieldgoals.xls. In this example we want to estimate the probability of kicking a field goal as a function of the distance (in yards) of kick.

(a) Fit the model to including distance and league. Examine the Lack of Fit tests, what do you conclude? Do any observations stand out as being unusual?

(b) Fit a model that also includes league × distance. Compare the Lack of Fit tests – have they improved? Does the interaction term appear to be significant? How do the AICc values compare? What do you recommend?