Project II: extension of due data

- DUE by Monday, Dec. 13 by 11am
- Location: 412 Thomas Building (my office, there will be an envelope/box marked for drop off).
- (1) You MUST send your data file by Wed. Dec. 8, 2004 via email to TA and me.
- (2) You MUST turn in TWO HARD copies of your report.
- (3) If you are welcome to turn in the project earlier. If you do so, but wish to submit a newer version by the final deadline, make sure that you clearly mark the most recent version.
- (4) IMPORTANT: I will NOT accept any late projects. Deadline is 11am!

At 11:10am the projects will be collected and by 11:30am you will get an email notifying you IF I DO NOT have a copy of your project indicating that you will receive zero points (so please do NOT wait until 10:55am to print the final version as something always goes wrong the last minute :-(). – so please ahead!

- (5) I hope to have project graded and final grades assigned by Wed. Dec. 15.

This Lecture

- Review:
  - Model Fit
  - Significance of the coefficients
  - Model Selection
- Prediction

- HW10 back
- HW11 turn-in

Course Evaluation

Model fit

- Read notes in HW 11
- Chapters 20 and 21 textbook
- Lecture notes handouts from the course website

The deviance goodness-of-fit statistics

- For testing the overall fit (adequacy) of the model
- The deviance statistics has an approximate chi-square distribution with \( n-p \) degrees of freedom
  - Think of \( n \) is the number of cells in a table (or number of observations), and \( p \) the number of parameters in the model
- Null hypothesis: the model we are testing fits the data well
- Alternative hypothesis: a different/more structure is needed to adequately model the outcome
- Large p-value indicates that the model is adequate
The deviance goodness-of-fit statistics
- SAS: Regression/Logistic/Statistics/Goodness-of-fit
- If the proportions are too small (e.g., counts per group less than 5) this measure could be misleading
- CAUTION: when have continuous explanatory variables, the number of groups is very large (every unique value of the continuous variable will create a new cell) so the above condition is rarely met
- Then,
  - obtain Hosmer-Lemeshow Statistics (large p-value indicates good model)
  - Or take the difference of log likelihoods (-2 Log L in the output under BETA=0 in SAS) for two models. This difference is approximately chi-squared with degrees of freedom equal to the difference in the number of parameters of the two models.

Significance of coefficients
- For each single coefficient the software gives an estimate of the coefficient with its standard error and the p-value.
- Null hypothesis: there is no relationship between the outcome and the explanatory variable
- Alternative hypothesis: there is a strong relationship
- Low p-value indicates that the predictor is significant (keep it in the model)

Model comparison
- Nested models
  - Take a difference of the Deviances and the degrees of freedom
  - The new statistics also follows chi-square distribution
  - Low p-value indicates a significant difference between the two models; and typically want the simpler model
- Non-nested models
  - In SAS look at AIC for example
  - The lower value indicates a better model
- In SAS: Logistic/Model/Selection

Prediction/Classification Tests
- Handouts
- The purpose of prediction test/classification is to determine if a person/unit/object belongs to the group with a specific characteristic.
- Some applications:
  - Drug use
  - Exposure to a disease
  - Pre-employment polygraph testing
  - Survival or not

Prevalence
- Widespread or a dominance of persons with a specific characteristic in a tested population
- Example: prevalence of persons with a specific characteristic in a population of interest.
- For example, let D represent a class of people with a characteristic (or a disease)
- For example, let S denote a membership to the group D, and Š a non-membership, as indicated by the test result.
- π = P(D)

Accuracy
- Sensitivity
  - The probability that a person with the specific characteristic is correctly classified
  - η = P[S|D]
- Specificity
  - The probability that a person who does NOT have a specific characteristic is correctly classified
  - θ = P[S|D̅]
Predictive value of a positive test (PVP)
- The conditional probability that a person whom test indicates belongs to a certain class/group actually does.
- $P[D|S]$

Predictive value of a negative test (PVN)
- The conditional probability that a person whom test indicates does NOT belong to a certain group actually does NOT belong.
- $P[D|\not{S}]$

False positive
- Mistakenly classify someone as with the characteristic
- $P[D|S]$
- $1 - PVP = 1 - P[D|S]$

False negative
- Mistakenly identify someone without the characteristic
- $P[D|\not{S}]$
- $1 - PVN = 1 - P[D|\not{S}]$

ROC = Receiver Operating Characteristics
- Handout_Accuracy.pdf
- Handout_ROC_Titanic.doc

In SAS
- Logistic/Statistics/Classification Table
  - E.g. for a single table with cutoff probability at 0.5 enter: From 0.5 to 0.5
- Logistic/Plot/ROC curve
- Logistic/Prediction/Predict New Data

Commands in SAS
- To create contingency tables, calculate chi-square statistic, etc…
  - Statistics/Table Analysis
- To run the logistic regression
  - Statistics/Regression/Logistic

Lessons from the course
- Overview/summary Lecture22.pdf
- You’ve learned something
  - If you understand some basic principles/concepts of statistics (e.g. difference between sample and population, statistics and parameters, …)
  - If you understand the use of some methods we covered in class
  - If you understand that the data analysis/interpretation is done within a context of a problem(s)/question(s)
  - If you can pick up a book and learn partial (or fully) on your own how to apply a method not covered in class

Next Lecture
- Presentation by Deet and Bill
- Course wrap-up
- Quiz grades
- Project II questions/turn-in