Chapter 8

Logistic Regression

Linear Regression

- Continuous response y
- Predictors x_j
- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \varepsilon_i$
- $\varepsilon_i \sim N(0, \hat{\sigma}^2)$

Linear Regression

Alternately we can write:

- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \varepsilon_i$
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi}$
- $Y_i \sim N(\hat{y}_i, \hat{\sigma}^2)$

Linear Regression

- Fine for continuous responses
- Model will technically predict any real value
- May only practically give a range of real values
- e.g. if observed predictors can only be positive, that might limit the modeled response

Extending to Other Types of Data

- What if we're interested in probability for an event?
- Response probability limited to 0<=p<=1
- Counts not normally distributed
- Counts follow Bernoulli (probability of individual trials) or Binomial distribution (number of events out of **n** trials)

Logistic Regression

Extension of concept of linear regression

- $Y_i \sim Ber(\hat{\pi}_i)$ with $\hat{y}_i = \hat{\pi}_i$
- Or $Y_i \sim Bin(n_i, \hat{y}_i/n_i)$ with $\hat{y}_i/n_i = \hat{\pi}_i$
- $\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots$

proc logistic

- For logistic regression
- Similar to **proc reg**, **proc anova**, **proc glm**, etc.
- Allows continuous and categorical predictors
- For 0|1 response data:

  model y = predictors

- For events out of trials:

  model events/trials = predictors

Initial Example: GHQ Data

- General Health Questionnaire
- Variables:
  - **ghq** test score (continuous)
  - **sex** (categorical; will ignore for now)
  - **cases** (counts)
  - **noncases** (counts)
- Linear and logistic models for probability of being a case based on **ghq** score

Odds

For binary cases:

- p/(1-p)
- Ratio of probability event happens to probability it doesn’t happen
- Odds of 1 means equally likely to happen or not

Odds Ratio

- Compares odds for event under different conditions
- In rough math

$$\text{oddsratio}(\text{event}; c1, c2) = \frac{odds(event|c1)}{odds(event|c2)}$$

- Odds ratio of 1 means odds are the same

Expected Odds and Probabilities

Under logistic regression model:

- Expected odds:

$$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots\right)$$

- Expected probability:

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots)}$$

Estimated Odds Ratio

- Starting with:

$$\frac{\hat{\pi}}{1-\hat{\pi}} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots\right)$$

- Change of 1 in x_j multiplies odds by $\exp\left(\hat{\beta}_j\right)$
- Odds ratio is significant (with respect to a change of 1 in the predictor) if 1 is not in confidence interval for odds ratio

Example: Logistic GHQ Model

- Look at results for case probability as a function of **ghq**

Adding a Categorical Predictor

- Add the **sex** variable to the model
- Coding of categorical variables
  - 0/1 (this is **reference cell coding**)
  - -1/1 (**deviations from means coding**)
- **sex** variable will be coded -1/1 in this example

Exercise: GHQ Model with gender

- Look at a scatter plot like before, except grouping data by **sex**
- Add the **sex** variable to the logistic model
- Interpret results
- Add **param=ref** option to **class** statement to see results using reference coding

Exercise: ESR and Plasma Data

- Response: **esr** 0|1 indicator for healthy or unhealthy erythrocyte sedimentation rate(ESR)
- 0 is healthy and 1 is unhealthy
- Predictors: **fibrinogen** and **gamma** globulin plasma levels
- Want to model unhealthy ESR as a function of plasma levels

Exercise: Logistic ESR Models

- For 0|1 data, **proc logistic** models 0 case by default
- Add **desc** option to **proc** statement to model 1's instead (and model unhealthy rates in this particular example)

Exercise: Logistic w/o Interaction Term

- Model unhealthy **esr** as a function of **gamma** and **fibrinogen** levels
- Which terms seem significant?
- What does the model tell us about odds ratios for changes in plasma levels?
- Remove insignificant terms, refit and interpret

Exercise: Adding Interaction Term

- Model unhealthy **esr** as a function of **gamma**, **fibrinogen** and their product
- Which terms seem significant?
- Use backward elimination to remove insignificant terms and comment on results

Predictions and Classifications

- Output data sets can include estimated probabilities and classifications
- We will obtain those results for the plasma model to see:
  - Predicted probabilities of unhealthy ESR
  - Frequencies for correctly and incorrectly classified observations
- How good is the classification?

Exercise: Do-It-Yourself data set

- Sample is from employed males aged 18 to 67
- Four categorical predictors
- Counts of yes and no responses to question about home improvements:
  - In previous year, did they do home improvements themselves that they would have previously hired someone to do?

Exercise: Do-It-Yourself data set

- Explanatory variables:
  - **work**: Skilled, Unskilled, or Office
  - **tenure**: rent or own
  - **type**: house or flat (apartment)
  - **agegrp**: 1, 2, or 3 (youngest to oldest)
- Response:
  - **yes**|**no** counts

Exercise: Do-It-Yourself data set

- Read in data and look at cross-classified probabilities of responding yes
- Note that we have categorical variables with more than two categories
- **param=ref ref=first** options in **class** statement will make first (in numeric or alphabetic order) the reference category

Exercise: **work** Only Model

- Model yes counts (probabilities) using only the **work** variable as a predictor
- Note the coding of variables
- What do we conclude about impact of type of work on this probability?
- What can we conclude about significant differences between the working classes?

Exercise: All Four Main Effects

- Fit the model with all four main effects (in the order given in the data set)
- What terms do and don't seem significant?
- What odds ratios do and don't seem significant?
- What conclusions can we draw from this model?

Exercise: Backward Selection

- Use backward selection to choose a best model
- What terms are significant?
- What odds ratios do and don’t seem significant?
- What conclusions can we draw from this model?