Building and Applying Logistic Regression Models
(Chapter 6)

MODEL SELECTION

Competing goals:

- Should be complex enough to fit the data well.
- Should be simple to interpret – should smooth the data rather than overfitting it.

Issue: How to select a parsimonious (simple) model that fits the data well?

- Unrealistic to hope to find the *true* model for a real dataset.
- Part science, part statistics, part experience and part common sense.
- Less number of parameters leads to more precise estimates.
- Watch out for *collinearity* – correlation in the estimated coefficients. If two covariates are highly correlated, do not need both of them in the model.

Indications of collinearity:

- Large standard errors.
- Look at the correlation matrix of the estimated coefficients. In R, use `cor2cov(vcov(fit))`, where `fit` contains the `glm` fit.

Indications of numerical instability:

- Error messages from the fitting program.
- Collinearity.
- Large standard errors.
- Zero or near-zero cell counts.
- Complete or near-complete separation. Complete separation means all zero responses appear at one combination of covariates and all one responses appear at another combination. No overlap in the covariates for the two responses. MLE does not exist in this case.

Models building strategy:

Step 1: Use univariate analysis to identify important covariates – the ones that are at least moderately associated with response.

- One covariate at a time.

- Analyze contingency tables for each categorical covariate. Pay particular attention to cells with low counts. May need to collapse categories in a sensible fashion.

- Use nonparametric smoothing for each continuous covariate. Can also categorize the covariate and look at the plot of mean response (estimate of π) in each group against the group mid-point. To get a plot on logit scale, plot the logit transformation of this mean response. This plot also suggests the appropriate scale of the variable.

- Can also fit logistic regression models with one covariate at a time and analyze the fits. In particular, look at the estimated coefficients, their standard errors and the likelihood ratio test for the significance of the coefficient.

- Rule of thumb: select all the variables whose p-value < 0.25 along with the variables of known clinical importance.

Step 2: Fit a multiple logistic regression model using the variables selected in step 1.

- Verify the importance of each variable in this multiple model using Wald statistic.

- Compare the coefficients of the each variable with the coefficient from the model containing only that variable.

- Eliminate any variable that doesn't appear to be important, and fit a new model. Check if the new model is significantly different from the old model. If it is, then the deleted variable was important.

- Repeat this process of deleting, refitting and verifying until it appears that all the important variables are included in the model.

- At this point, add the variables into the model that were not selected in the original multiple model. Assess the joint significance of the variables that were not selected. This step is important as it helps to identify the confounding variables. Make changes in the model, if necessary.

At the end, we have the *preliminary main effects model* – it contains the important variables.

Step 3: Check the assumption of linearity in logit for each continuous covariate.

- Look at the smoothed plot of logit in step 1 against the covariate.
- If not linear, find a suitable transformation of the covariate so that the logit is roughly linear in the new variable.
- Try simple transformations such as power, log, etc. Also read about the method of fractional polynomials in the handout.

At the end, we have the *main effects model*.

Step 4: Check for interactions.

- Create a list of possible pairs of variables in the main effects model that have some scientific basis to interact with each other. This list may or may not consist of all possible pairs.

- Add the interaction term, one at a time, in the model containing all the main effects and assess its significance using the likelihood ratio test.

- Identify the significant interaction terms.

Step 5: Add the interactions found significant in step 4 to the main effects model and evaluate its fit.

- Look at the Wald tests and LR tests for the interaction terms.
- Drop any non-significant interaction.

At the end, we get our *preliminary final model*. We should now assess the overall goodness-of-fit of this model and perform model diagnostics.

Example: Read the analysis of UMARU impact study in the handout. You are expected to do the analysis for your project on similar lines.

Another strategy: Automatic stepwise selection procedure.

- Start with a list of important covariates obtained as before using the univariate analysis.
- *Forward selection*: Start with a simple model and add terms sequentially until further additions do not significantly improve the fit.
- *Backward elimination*: Start with a complex model and remove terms sequentially until a further deletion leads to a significantly poorer fit. (Generally preferred over forward selection).
- Other variants.
- Cannot trust the results.
- Can also use a penalized measure of model fit such as Akaike Information Criterion (AIC) instead of p-values. AIC = – 2(maximized log likelihood – # parameters in the model). Lower is better.

Example: Read section 6.1.3 for an example and Laura Thompson's manual for R code.

DETECTING LACK OF FIT

We have a model that we are reasonably satisfied with. The model fits well if the observed and fitted responses are close. With categorical covariates, it is likely that their # of distinct settings is (much) less than N. In other words, several subjects may have the same covariate setting.

Goal: Identify covariate patterns with lack of fit.

- J = # distinct covariate settings (patterns).
- m_j = # of subjects in the j-th pattern. Then $m_1 + m_2 + \ldots + m_J =$ ___.
- o_j = # of observed successes in the j-th covariate pattern.
- e_j = fitted # of successes in the j-th covariate pattern = _____.
- Plot the observed versus fitted counts. If model fits well, the points should be close to the 45^o line through origin. This method is effective only when J _____ n
- With continuous covariates, it is likely that J ___ n.

Example: Suppose there are 25 subjects in a study with 3 covariates – SEX, RACE and WEIGHT. We have 12 Males, 13 Females, 10 Whites, 8 Blacks and 7 Hispanics. Further, no two have the same weight.

1. Model has only SEX. Then J = _____, m_1 = _____, m_2 = _____.
2. Model has both SEX and RACE. Then J = _____.
3. Model has all three covariates. Then J = _____.

- Construct the following 2 x J contingency table and analyze the Pearson and deviance residuals for the observed and the fitted counts in this table.

Covariate Pattern

	1	2	…	J	Total
Observed # y = 1 Fitted # y = 1	o_1 $e_1 = m_1\hat{\pi}_1$	o_2 $e_2 = m_2\hat{\pi}_2$		o_J $e_J = m_J\hat{\pi}_J$	
Observed # y = 0 Fitted # y = 0	$m_1 - o_1$ $m_1 - e_1$	$m_2 - o_2$ $m_2 - e_2$		$m_J - o_J$ $m_J - e_J$	
Total	m_1	m_2		m_J	N

1. Pearson residual: $r_j = \dfrac{o_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$

2. Deviance residual: $d_j = \pm 2\left[o_j \ln\left(\dfrac{o_j}{e_j}\right) + (m_j - o_j)\ln\left(\dfrac{m_j - o_j}{m_j - e_j}\right)\right]$, where the sign +/- is same as the sign of $(o_j - e_j)$.

- Both residuals are close to zero the observed and fitted counts are close.

- **Recall**: Roughly speaking, when m_j is large and the fitted model is correct, r_j and d_j are approximately normal with mean zero and variance less than one. (Their standardized versions have variance one). Their absolute values larger than 2 or 3 indicate a possible lack of fit.

- Rule of thumb for normal approximation: Almost every cell should have a fitted count at least 5.

INFLUENCE DIAGNOSTICS

Goal: Identify observations that have too much influence on fitted model.

- Delete one observation at a time, and look at the change in fit of the model and the estimates.
- Observation could be a single individual or a covariate pattern.
- Case-deletion diagnostics.

Popular measures:

- Dfbetas = change in the estimated coefficients divided by its SE.
- Dffits = change in the fitted value divided by its SE.
- Cook's distance = standardized sum of squares of change in all fitted values.
- covratio = change in covariance matrix of the estimates.
- Change in the Pearson or the LR χ^2 statistic.
- Most of these can be obtained in R using `influence.measures(fit),` where `fit` contains the `glm` fit.

Using the diagnostics:

- Plot them against the estimated probabilities.
- Look for outlying points.

What to do?

- Do not expect to identify many poorly fit or influential points when the model seems to fit well on overall goodness-of-fit measures (e.g., Hosmer-Lemeshow test).

- When there many such observations, one or more of following happened:

    - the logistic model is not a good approximation to the true relationship between $\pi(\underline{x})$ and $\underline{x}$.
    - an important covariate is missing
    - at least one of the covariates doesn't have its correct scale in the model.

- Sometimes these problems can be alleviated by going back to the model building step.