Review
Chapter 4 parts 3,4 and 5

(and Chapters 1, 2, 3 & 4 parts 1,2)

ERHS 642

Spring 2014

Logistic Regression

- Build a model
  - to predict or explain the probability of a dichotomous outcome (0/1)
  - given a particular constellation of risk factors
- Probability is S shaped
- The logistic regression model is a good choice because it provides adjusted ORs

Maximum likelihood estimation

- Create likelihood function expressing the probability of actually observing the data we observed
- Choose coefficients β_0 and β_1 such that the likelihood function is maximized
- Given $\hat{\beta}_0$ and $\hat{\beta}_1$, the outcome predicted by the model mirrors the observed outcome most closely

Significance tests

- The Wald test and the Likelihood Ratio test can be used to test if the model with certain variables is better than the model without these variables
- The Likelihood Ratio test is "better" than the Wald test
- The more variables are in a logistic regression model, the lower the power

Logit differences / OR estimation

To estimate an OR of interest,

- Determine the logit
- Determine the values used to calculate the logit difference
- Calculate the logit difference
- Exponentiate the result

Confounding in logistic regression

x = risk factor c = confounder

- Run the model containing x only, estimate the OR for x (crude OR)
- Run the model containing x and c, estimate the OR for x (adjusted OR)
- Use the "10% rule" to compare the ORs

Multiplicative interactions in logistic regression

x = risk factor c = potential effect modifier

To determine if x and c interact,

- Run the logistic regression model containing x, c and x×c
- If x*c is statistically significant at the 0.1 level
  - Calculate the appropriate logit differences and use contrast statement to calculate ORs
- If x×c is not statistically significant at the 0.1 level, remove the interaction term from the model

Additive interactions in logistic regression

x = risk factor c = potential effect modifier

To determine if x and c interact,

- Run the linear link regression model containing x, c and x×c (check if $0 \leq \hat{\pi} \leq 1$)
- If x×c is statistically significant at the 0.1 level
  - Calculate the appropriate logit differences and use contrast statement to calculate ORs for a 4-row table
- Or use a 4-row table

Why do we assess the scale of a continuous covariate?

- Continuous model covariates are assumed to increase linearly in the logit
- Example: x=age, y=adverse birth outcome
  - The model implicitly assumes that an increase in age of, say, 5 years has the same effect on the logit no matter what age we start at
  - This doesn't make sense biologically

How do we assess the scale of a continuous covariate?

- Splines
- Categorizing
- Fractional polynomials

Spline method

- Select "knots", i.e. connection points
  - Choose number of knots and spacing
- Select connections between knots
  - Constant, linear, cubic, other connection
- Determine non-linearity in the logit
  - Visual assesment
  - Check for stat. significance

Pros and cons of the spline method

- PROs:
  - Easy to use
  - Quick method to check for non-linearity in the logit
  - Can compare different splines and select "the best" model, i.e. the model with the smallest deviance
- CONs:
  - Too many choices (knots, connections)
  - Still, many possible knots and connections are not tested
  - Must check for statistical significance! The shape of the plot may just be noise.
  - Very difficult to obtain interpretable ORs; therefore, finding the best spline model is not very useful in practice

Categorizing

- Based on the spline plots it may be possible to establish cutpoints
- Alternatively, quartiles can be used
- Biologically meaningful cutpoints can also be used
- For resulting categorical variables with more than 2 categories, design variables must be used in the model

Pros and cons of categorizing

- PROs:
  - Easy to do
  - If meaningful cutpoints are chosen, results and conclusions are meaningful
- CONs:
  - Increases the number of variables in the model
  - May result in misclassification

fp method

- Model the continuous variable using many different scales (e.g., linear, quadratic, cubic, log-transformed, etc.)
- Compare the different models and select "the best" model, i.e. the model with the smallest deviance (consider statistical significance!)
- Transform the continuous variable accordingly

Pros and cons of the fp method

- PROs:
  - Automated
  - Lots of possible transformations tested
- CONs
  - Best transformation may be very complex and hard to interpret/explain to lay persons
  - Based on statistical significance
  - Many possible transformations are not tested
  - Values of the variable of interest must be > 0 (transformations include division by the variable and the natural log (ln) of the variable)

:

Modeling the variable "glasses of alcohol per day"

- Dichotomize (0=non-drinker and 1=drinker)
- Use dichotomous and continuous variable
- Estimate OR (drinking x+c glasses per day vs. drinking x glasses per day)
- Estimate OR (drinking c glasses per day vs. not drinking)

Or:

- Categorize the variable

Possible reasons for zero cells

- Random error (sample size too small)
- Systematic error
- True absence of subjects in the category
- Complete separation

How do zero cells affect a logistic regression analysis?

- The model falls apart

Complete separation

- One or more covariates perfectly predict the outcome

Possible reasons for complete separation

- Random error
- Systematic error
- True absence of subjects in the categories
- Overfitting the model

How does complete separation affect a logistic regression analysis?

- Creates zero cells
- The model falls apart
- The variable cannot be used

Quasi-complete separation

- One or more covariates <u>almost</u> perfectly predict the outcome

How does quasi-complete separation affect a logistic regression analysis?

- The model falls apart
- The variable cannot be used

Collinearity

- Two or more variables are identical

How does collinearity affect a logistic regression analysis?

- One of the variables is set to 0

What do we do in the presence of collinearity?

- Use only one of the variables

Name 3 potential goals of a logistic regression analysis

- Goal 1: To get the most complete "picture" of the risk factors for the outcome
- Goal 2: To get the most complete "picture" of one specific risk factor
- Goal 3: To best predict the outcome

What analyses should be conducted prior to model selection

- Get to know the study variables
  - Cross-tabulate categorical variables
  - Calculate descriptive statistics for continuous variables
  - Locate any unusual or incorrect values
- If necessary, make changes to the study variables
  - Delete or correct unusual or incorrect values
  - Collapse categories
  - Remove categories
  - Remove variables

Describe the approach to purposeful model selection for goal 1

- Statistically significant variables, confounders and effect modifiers should be included in the model
- Univariate significance ($p<0.25$)/biological importance
- Univariate scale
- Multivariate significance ($p<0.05$)/biological importance
- Multivariate scale
- Confounding (10% rule)
- Interactions ($p<0.1$)
- Model stability

Describe the approach to purposeful model selection for goal 2

- The risk factor and confounders and effect modifiers of the risk factor should be included in the model
- Other statistically significant variables may or may not be included
- Bi-/trivariate analyses (confounding / effect modification of the risk factor)
- Multivariate analyses (confounding / effect modification of the risk factor)

Describe the approach to purposeful model selection for goal 3

- Confounders and effect modifiers are only important if they improve the predictive ability of the model
- Analyses of predictive ability of the model

How many variables can be included in a logistic regression model?

- Rough guide:

  No more variables than the "least frequent outcome" divided by 10

What can we do when more variables than we should use are significant or important

- Use more variables than recommended but look out for model instability and wide CIs
- Concentrate on statistically or biologically important variables
- Reduce the number of confounders / effect modifiers

What is the idea behind stepwise selection?

At each step,

- Test significance of variables when added to the model
- Add most significant variable (if $p < p_{Entry}$)
- Remove model covariates with $p > p_{Exit}$
- Stop when no more variables have $p < p_{Entry}$

What is the idea behind best subsets selection?

- Model all combinations of 2, 3, 4, etc. variables and compare the resulting models to the model containing all independent variables
- In theory: Use Mallow's C_q to decide which models are best
- In this class: Look for confounders you may have missed

What are the advantages and disadvantages of stepwise selection?

Pros

- Quick and easy (kind of…not really…)

Cons

- Confounders may be missed
- Biological/clinical importance is ignored
- Model stability is ignored
- Design variables

What are the advantages and disadvantages of best subsets selection?

Pros

- Quick and easy (sort of…)
- May find model covariates you may otherwise miss

Cons

- Biological/clinical importance is ignored
- Model stability is ignored
- Design variables

How should we treat design variables in stepwise selection?

- Could create design variables in data step
  - Must decide what to do if only part of the set of design variables is included
- Could use class statement
  - OK if the variable is included
  - If the variable is not included, we don't know if part of the set of design variables is significant

How should we treat design variables in best subsets selection?

- Must create design variables in data step
  - Must decide what to do if only part of the set of design variables is included
- The class statement doesn't work with best subsets selection

Can we use automated selection procedures for interactions?

- Yes
- Important to keep in mind that quasi-complete or complete separation may prematurely stop the selection procedure