Lecture 12
Model Building

BMTRY 701
Biostatistical Methods II

The real world regression

- datasets will have a large number of covariates!
- There will be a number of covariates to consider for inclusion in the model
- The inclusion/exclusion of covariates
  - will not always be obvious
  - will be affected by multicollinearity
  - will depend on the questions of interest
  - will depend on the scientific ‘precedents’ in that area
- The model building process is important for determining a “final model”

The “final model”

- At the end of the analytic process, there is generally one model from which you make inferences
- it usually is a multiple regression model
- it is not logical to make inferences based on more than one model
- Recall the ‘principle of parsimony’

Principle of Parsimony

- Also known as Occam's Razor
- The principle states that the explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis or theory.
- The principle recommends selecting the hypothesis that introduces the fewest assumptions and postulates the fewest entities
- Translation for regression:
  - the fewest possible covariates that explains the greatest variance is best!
  - The addition of each covariate should be weighed against the increase in complexity of the model.

General Process of Model Building

1. Exploratory Data analysis
2. Choose initial model
3. Fit model
4. Check model assumptions
5. Repeat 2 – 4 as needed
6. Interpret findings

Exploratory Data Analysis

- Consider the covariates and the outcome variables
  - look at each covariate and outcome
    - what forms do they take?
    - might transformations need to be made?
  - look at relationships between Y and each X
    - are the relationships linear?
    - what form should a covariate take to enter the model (e.g. categorical? spline? quadratic?)
  - look at the relatioships between the X's
    - is there strong correlation between some covariates?

Exploratory Data Analysis

- Individual variable analysis
  - histograms
  - boxplots
  - dotplots (by categories?)
- Two-way associations
  - scatterplots
  - color-coded by third variable?
  - SIMPLE LINEAR REGRESSIONS
- For categorical variables
  - tables
  - color code other graphical displays

SENIC

ID	ID Number	1-113
LOS	Length of stay	Average length of stay of all Patients in hospital (in days)
AGE	Age	Average age of patients (in years)
INFRISK	Risk of infection	Average estimated probability of Acquiring infection in hospital (in percent)
CULT	Routine culturing ratio	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100
XRAY	Routine chest X-ray ratio	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100
BEDS	Number of beds	Average number of beds in hospital in study period
MEDSCHL	Medical school affiliation	1 = Yes, 2 = No
REGION	Region	Geographic region, where: 1 = NE, 2 = NC, 3 = S, 4 = W
CENSUS	Average daily census	Average number of patients in hospital per day during study period
NURSE	Number of nurses	Average number of full-time equivalent registered and licensed practical nurses during study period (number full-time plus one half the the number part-time)
FACS	Available facilities and services	Percent of 35 potential facilities and services that are provided by the hospital

SENIC example

- We need a scientific question/hypothesis!!
- Examples:
  - What factors are predictive of length of stay?
  - Is the number of beds strongly related to length of stay?
  - Is there a difference in length of stay by region?
  - how do infection risk and number of cultures relate to length of stay? is it possible to reduce the length of stay by reducing infection risk and number of cultures?

Work through the exploration…

Next step: Pick an initial model

- Use the information that you learned in the exploratory step
- Some guidelines
  - covariates not associated in SLR models will probably not be associated in MLR models
  - Choose threshold: alpha < 0.10 or 0.20 in SLR to be included in initial MLR
- Recall multicollinearity
  - might want to spend some extra time learning about the interrelationships between two variables and the outcome.

Next step: Pick an initial model

- Many approaches to the initial model
- My approach: start big, and then pare down
  - initial model includes all of the covariates and potentially their interactions
  - fit model with all of the covariates of interest
  - remove ONE AT A TIME based on insignificant p-values and model coefficients
    - find the most insignificant covariate
    - refit the model without it
    - look at model:
      - what happened to other coefficients?
      - what happened to R^2
  - not hard-fast rules!

SENIC

- What is an appropriate initial model?
- Are there any interactions to consider?
- Work through the model…

Check model assumptions

- Based on a reasonable model (in terms of ‘significance’ of covariates), check the assumptions
- Residual plots
- Other diagnostics
- Recall your assumptions:
  - independence of errors
  - homoscedasticity/constant variance
  - normality of errors

Does it fit?

- If so….go to next step
- If not, deal with misspecifications
  - transform Y?
  - another type of regression?!
  - transform X?
  - consider more exploration (e.g., smoothers to inform about relationships)
  - outlier problems?
  - Then, refit all over again…

Last step

- Interpret results
- Oddly, this step often leads you back to refitting
- Sometimes trying to summarize results causes you to think of additional modeling considerations
  - adding another variable
  - using a different parameterization
  - using a different reference level for a categorical variable

SENIC

- What is the final model?
- How to present it?

Other model building issues: Stepwise approaches

- “Stepwise” approaches are computer driven
- you give the computer a set of covariates and it finds an ‘optimal’ model
- “forward” and “backward”
- Problems:
  - models are only ‘stepwise’ optimal
  - ignore magnitude of β and simply focus on p-value!
  - you need to set criteria for optimality which are not always obvious
  - gives you no ability to give different variables different priorities
  - can have problematic interpretations: e.g. a main effect is removed, but the interaction is included.
  - stepwise forward and backward give different models.

Is stepwise ever a good idea?

- If you have a very large set of predictors that are somewhat ‘interchangeable’
- Example: gene expression microarrays
  - you may have >10000 genes to select from
  - automated procedures can find optimal set that describe a large amount of variation in the outcome of interest (e.g. cancer vs. no cancer)
  - it would be physically impossible to use manual model-fitting
  - Specialized software for this (standard ‘lm’ type approach will not work).

Stepwise Approaches

- I don’t condone it but,
- In R: step(reg)

Other model building issues: R^2

- Some people use increase in as a criteria of inclusion/exclusion of a covariate
- Not that common in biomedical research, but not totally absent either
- Look at either
  - coefficient of *multiple* determination
  - coefficient of *partial* determination
- Tells the fraction of variance explained.

Other model building issues: Information Criteria

- Information Criteria (IC)
  - help with choosing between two models
  - compare ‘parsimony’ adjusted statistics.
  - choose the model with the smallest IC
  - AIC = Akaike information criteria
  - BIC = Bayesain information criteria
- More with logistic regression…

A step further

- Model validation
- Addresses issue of ‘overfitting’ etc.
- Is this model specific to this dataset, or does it actually work in the general setting?
- Need to
  - collect more data
  - split data into ‘training set’ and ‘test set’
  - other cross-validation approaches such a ‘leave-k out’ approach

Next: Diagnostics in MLR

- Added variable plots
- Identifying outliers
  - hat matrix: shows leverage and influence
  - studentized or standardized residuals
- variance inflation factor