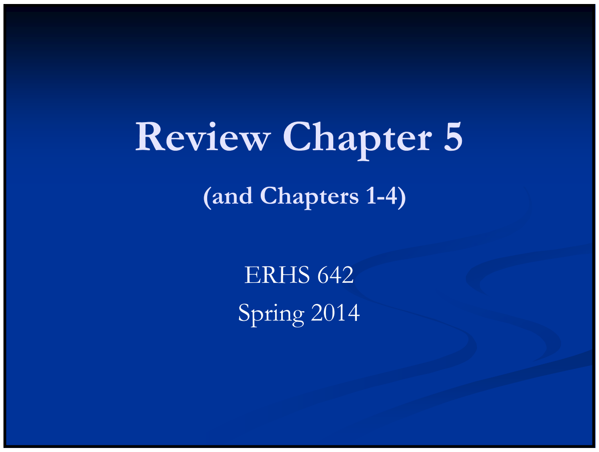
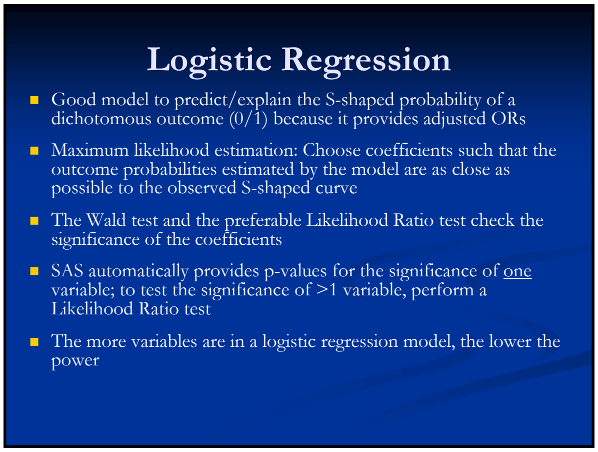


Review Chapter 5

(and Chapters 1-4)

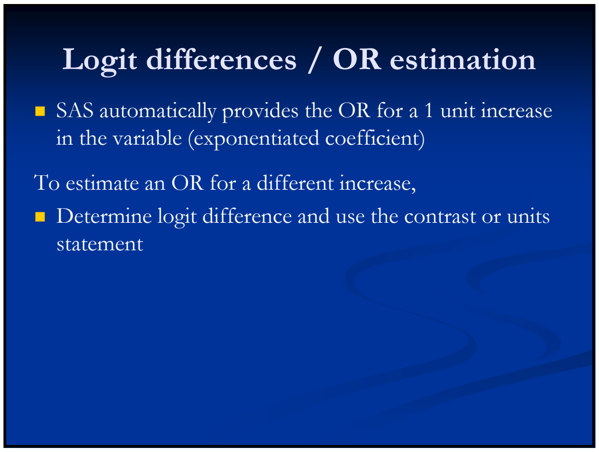
ERHS 642

Spring 2014



Logistic Regression

- Good model to predict/explain the S-shaped probability of a dichotomous outcome (0/1) because it provides adjusted ORs
- Maximum likelihood estimation: Choose coefficients such that the outcome probabilities estimated by the model are as close as possible to the observed S-shaped curve
- The Wald test and the preferable Likelihood Ratio test check the significance of the coefficients
- SAS automatically provides p-values for the significance of one variable; to test the significance of >1 variable, perform a Likelihood Ratio test
- The more variables are in a logistic regression model, the lower the power

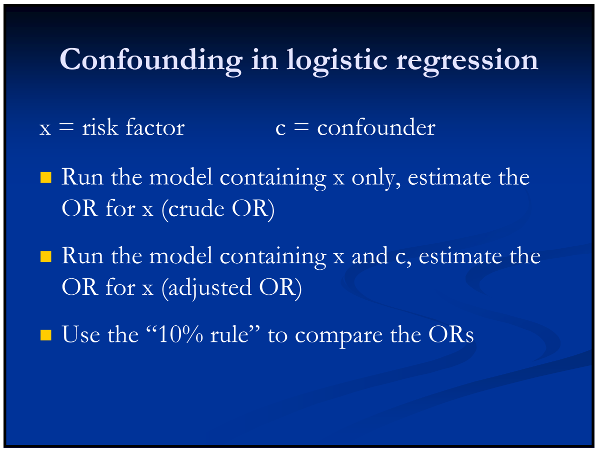


Logit differences / OR estimation

- SAS automatically provides the OR for a 1 unit increase in the variable (exponentiated coefficient)

To estimate an OR for a different increase,

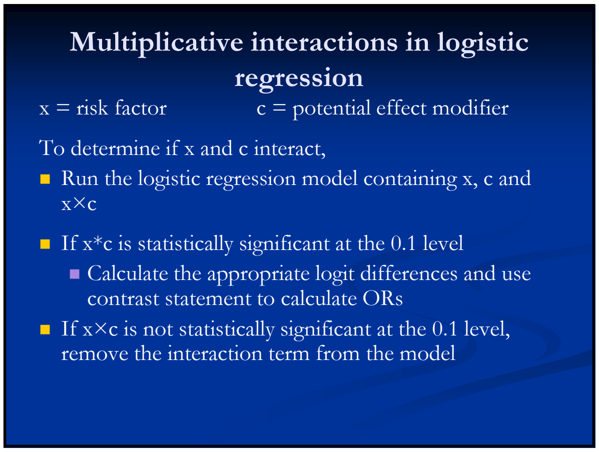
- Determine logit difference and use the contrast or units statement



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

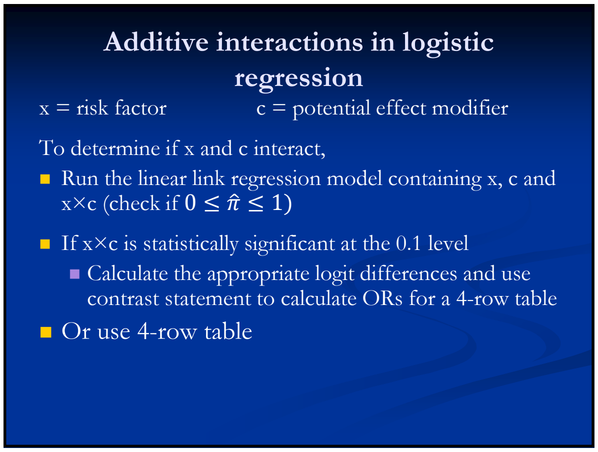


Multiplicative interactions in logistic regression

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

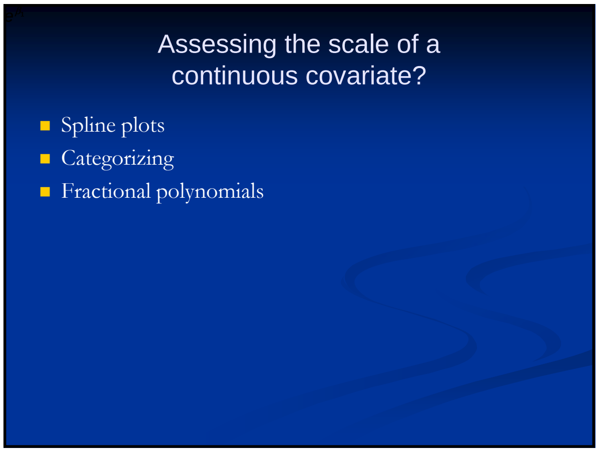


Additive interactions in logistic regression

x = risk factor c = potential effect modifier

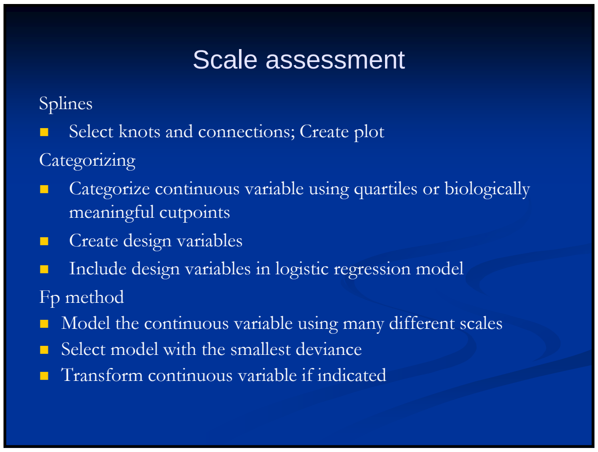
To determine if x and c interact,

- Run the linear link regression model containing x, c and x×c (check if $0 \le \hat{\pi} \le 1$)
- 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 4-row table



Assessing the scale of a continuous covariate?

- Spline plots
- Categorizing
- Fractional polynomials



Scale assessment

Splines

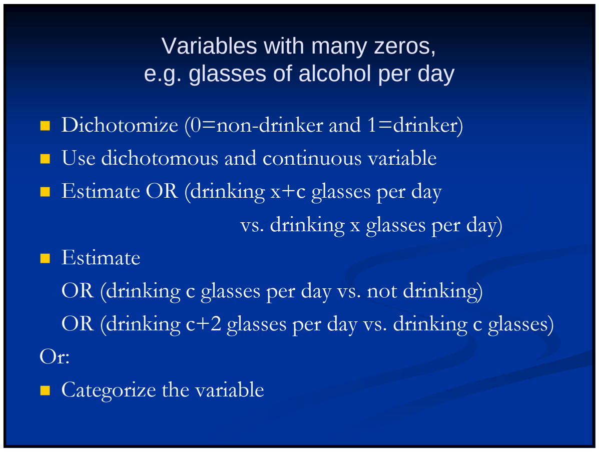
- Select knots and connections; Create plot

Categorizing

- Categorize continuous variable using quartiles or biologically meaningful cutpoints
- Create design variables
- Include design variables in logistic regression model

Fp method

- Model the continuous variable using many different scales
- Select model with the smallest deviance
- Transform continuous variable if indicated

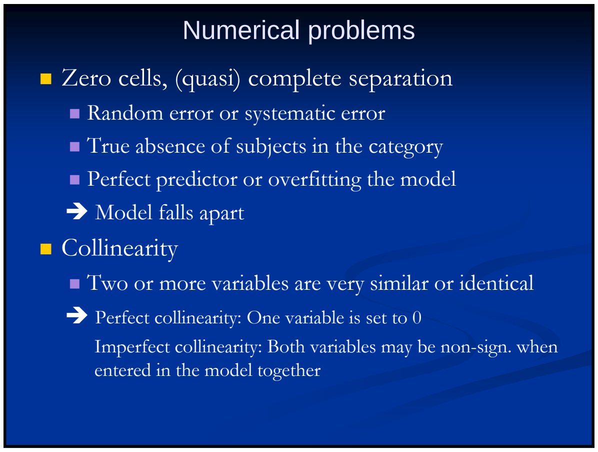


Variables with many zeros, e.g. glasses of alcohol per day

- 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 (drinking c+2 glasses per day vs. drinking c glasses)

Or:

- Categorize the variable



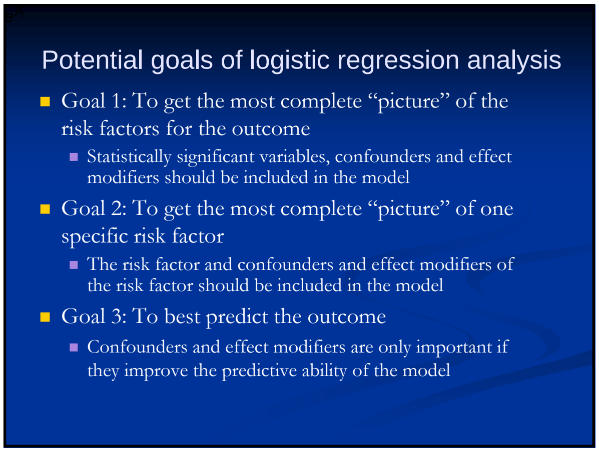
Numerical problems

- Zero cells, (quasi) complete separation
  - Random error or systematic error
  - True absence of subjects in the category
  - Perfect predictor or overfitting the model

  ➔ Model falls apart
- Collinearity
  - Two or more variables are very similar or identical

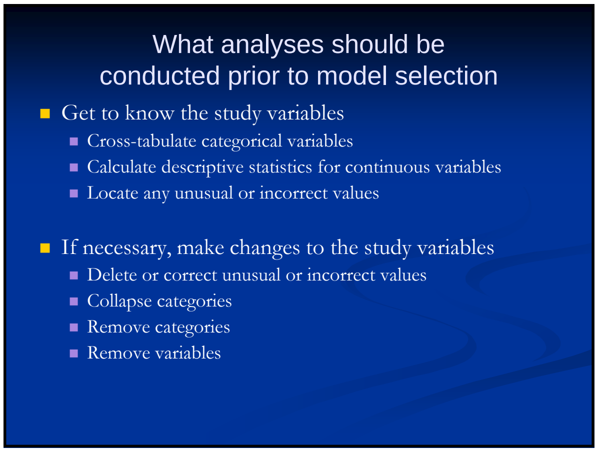
  ➔ Perfect collinearity: One variable is set to 0

  Imperfect collinearity: Both variables may be non-sign. when entered in the model together



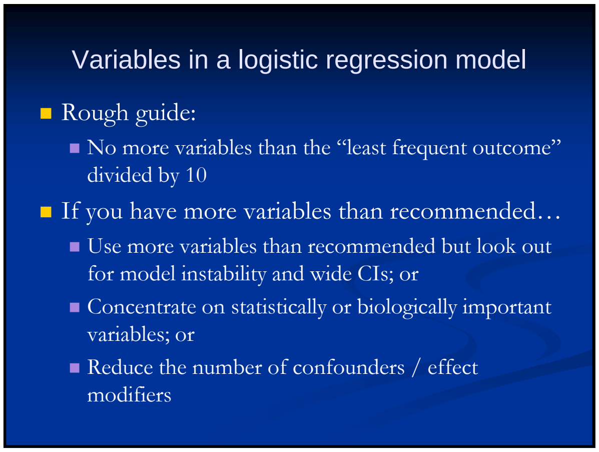
Potential goals of logistic regression analysis

- Goal 1: To get the most complete "picture" of the risk factors for the outcome
  - Statistically significant variables, confounders and effect modifiers should be included in the model
- Goal 2: To get the most complete "picture" of one specific risk factor
  - The risk factor and confounders and effect modifiers of the risk factor should be included in the model
- Goal 3: To best predict the outcome
  - Confounders and effect modifiers are only important if they improve the predictive ability of the model



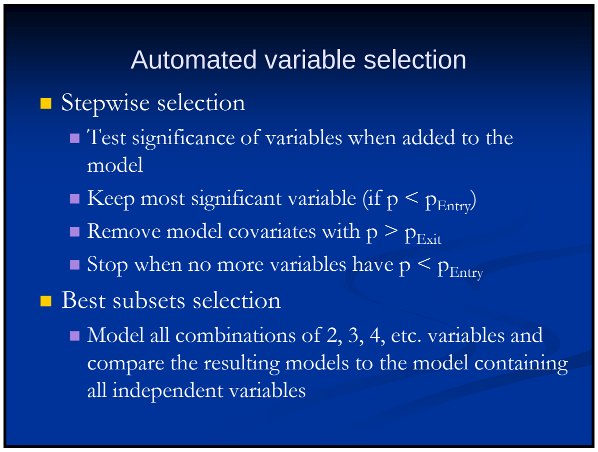
What analyses should be conducted prior to model selection

- 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



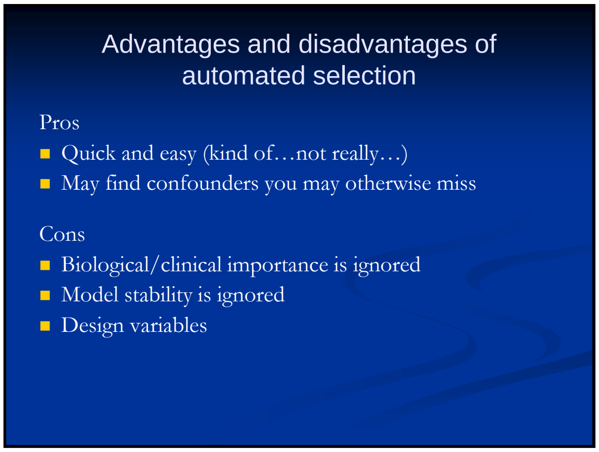
Variables in a logistic regression model

- Rough guide:
  - No more variables than the "least frequent outcome" divided by 10
- If you have more variables than recommended…
  - Use more variables than recommended but look out for model instability and wide CIs; or
  - Concentrate on statistically or biologically important variables; or
  - Reduce the number of confounders / effect modifiers



Automated variable selection

- Stepwise selection
  - Test significance of variables when added to the model
  - Keep most significant variable (if $p < p_{Entry}$)
  - Remove model covariates with $p > p_{Exit}$
  - Stop when no more variables have $p < p_{Entry}$
- Best subsets selection
  - Model all combinations of 2, 3, 4, etc. variables and compare the resulting models to the model containing all independent variables



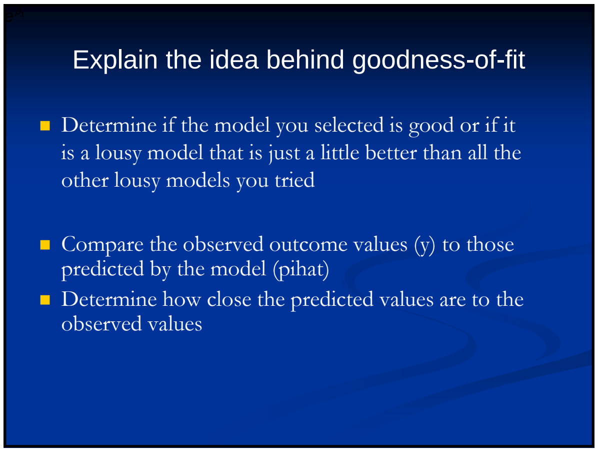
Advantages and disadvantages of automated selection

Pros

- Quick and easy (kind of…not really…)
- May find confounders you may otherwise miss

Cons

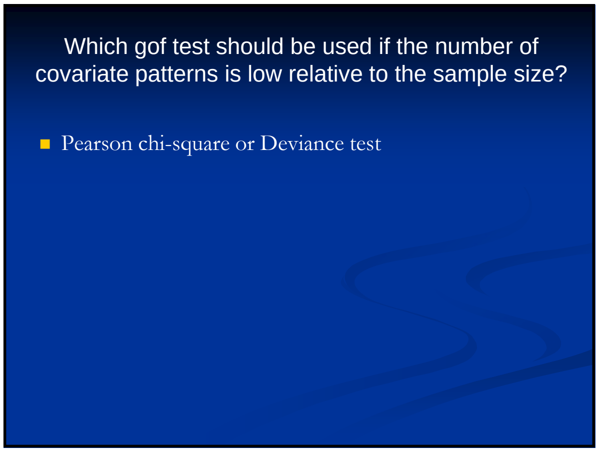
- Biological/clinical importance is ignored
- Model stability is ignored
- Design variables



Explain the idea behind goodness-of-fit

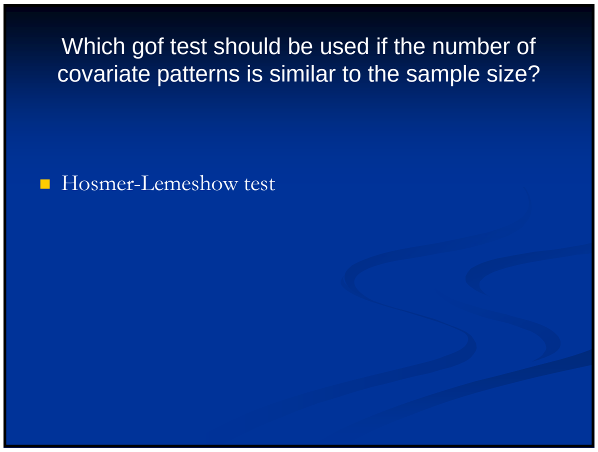
- Determine if the model you selected is good or if it is a lousy model that is just a little better than all the other lousy models you tried

- Compare the observed outcome values (y) to those predicted by the model (pihat)
- Determine how close the predicted values are to the observed values



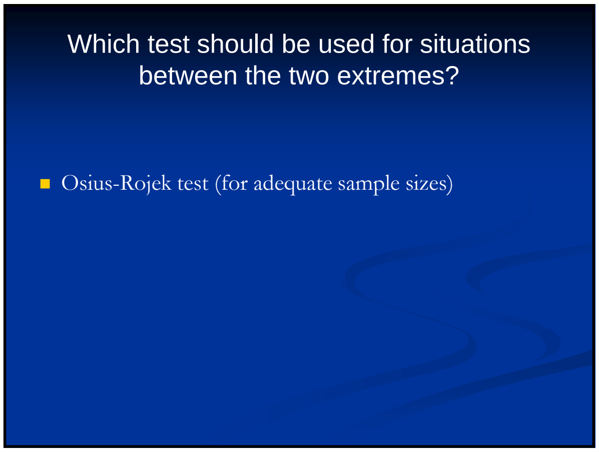
Which gof test should be used if the number of covariate patterns is low relative to the sample size?

- Pearson chi-square or Deviance test



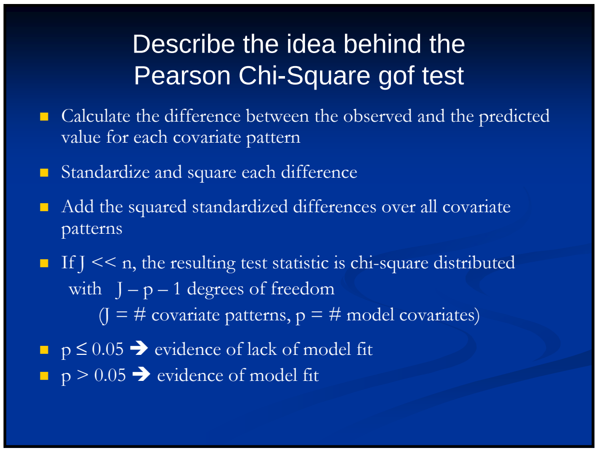
Which gof test should be used if the number of covariate patterns is similar to the sample size?

- Hosmer-Lemeshow test



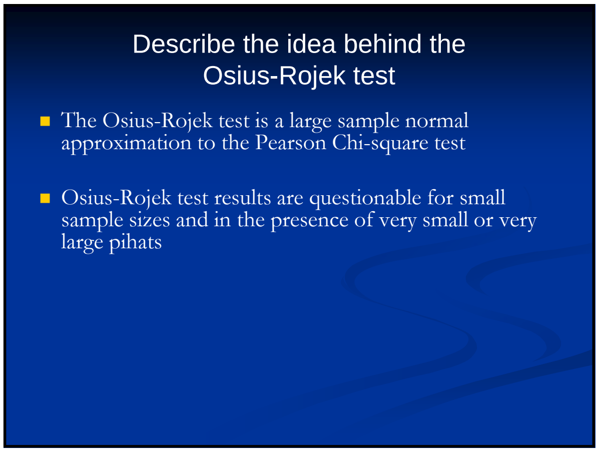
Which test should be used for situations between the two extremes?

- Osius-Rojek test (for adequate sample sizes)



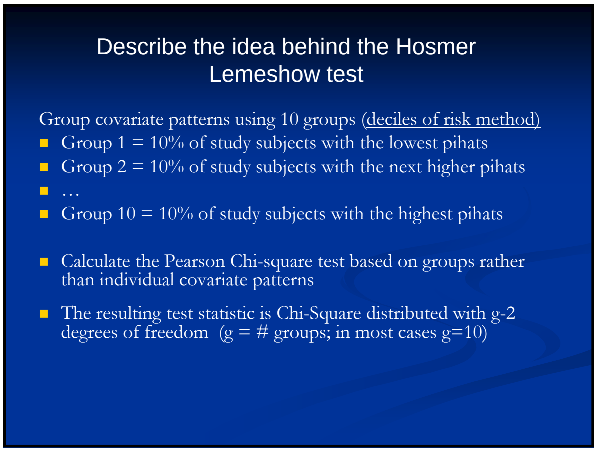
Describe the idea behind the Pearson Chi-Square gof test

- Calculate the difference between the observed and the predicted value for each covariate pattern
- Standardize and square each difference
- Add the squared standardized differences over all covariate patterns
- If J << n, the resulting test statistic is chi-square distributed
  with J – p – 1 degrees of freedom
  (J = # covariate patterns, p = # model covariates)
- p ≤ 0.05 ➔ evidence of lack of model fit
- p > 0.05 ➔ evidence of model fit



Describe the idea behind the Osius-Rojek test

- The Osius-Rojek test is a large sample normal approximation to the Pearson Chi-square test
- Osius-Rojek test results are questionable for small sample sizes and in the presence of very small or very large pihats

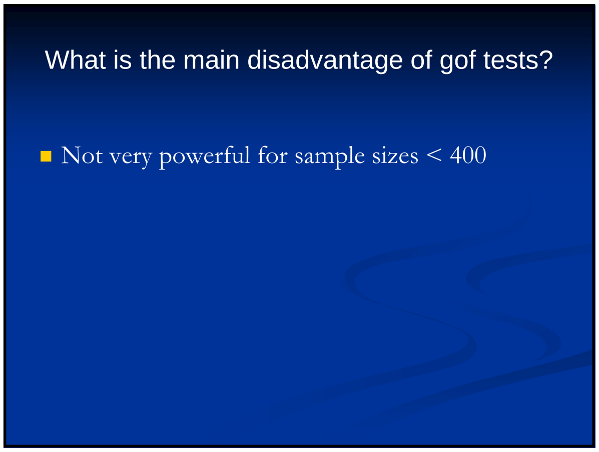


Describe the idea behind the Hosmer Lemeshow test

Group covariate patterns using 10 groups (deciles of risk method)

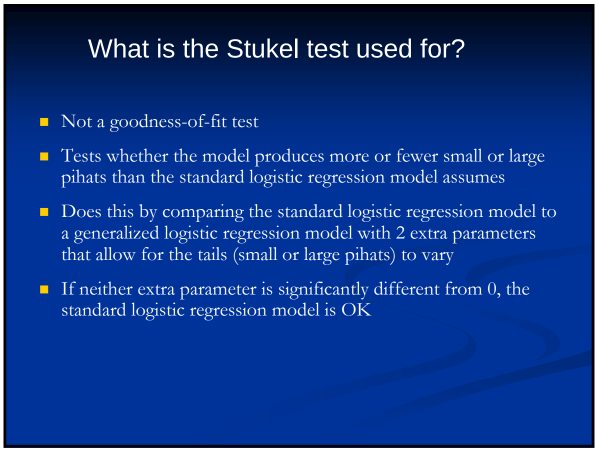
- Group 1 = 10% of study subjects with the lowest pihats
- Group 2 = 10% of study subjects with the next higher pihats
- …
- Group 10 = 10% of study subjects with the highest pihats

- Calculate the Pearson Chi-square test based on groups rather than individual covariate patterns
- The resulting test statistic is Chi-Square distributed with g-2 degrees of freedom (g = # groups; in most cases g=10)



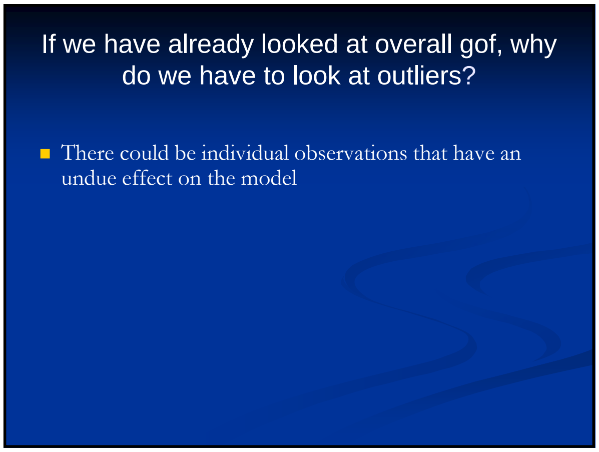
What is the main disadvantage of gof tests?

- Not very powerful for sample sizes < 400



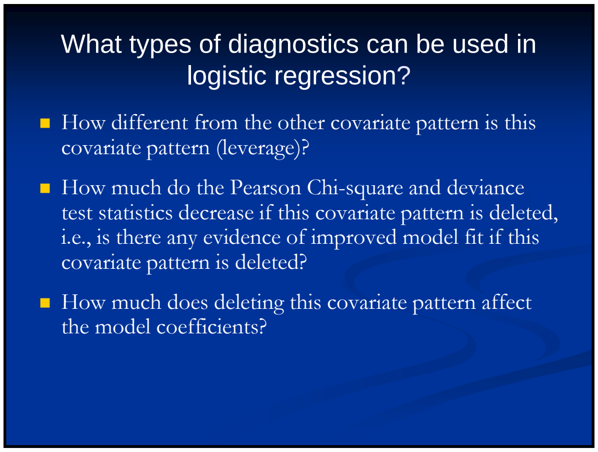
What is the Stukel test used for?

- Not a goodness-of-fit test
- Tests whether the model produces more or fewer small or large pihats than the standard logistic regression model assumes
- Does this by comparing the standard logistic regression model to a generalized logistic regression model with 2 extra parameters that allow for the tails (small or large pihats) to vary
- If neither extra parameter is significantly different from 0, the standard logistic regression model is OK



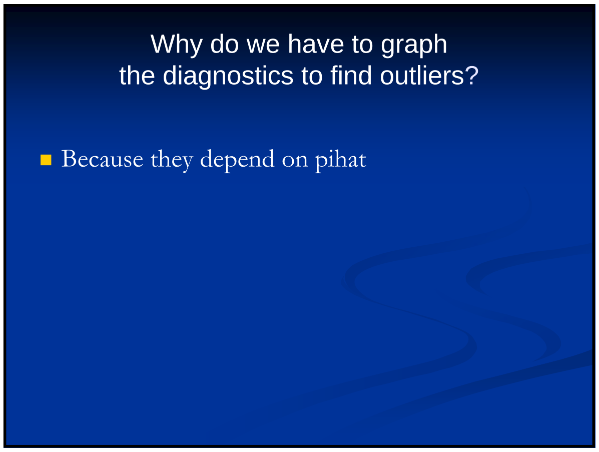
If we have already looked at overall gof, why do we have to look at outliers?

- There could be individual observations that have an undue effect on the model



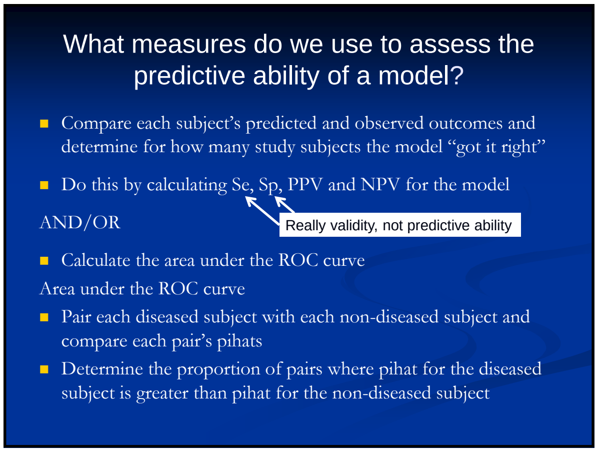
What types of diagnostics can be used in logistic regression?

- How different from the other covariate pattern is this covariate pattern (leverage)?
- How much do the Pearson Chi-square and deviance test statistics decrease if this covariate pattern is deleted, i.e., is there any evidence of improved model fit if this covariate pattern is deleted?
- How much does deleting this covariate pattern affect the model coefficients?



Why do we have to graph the diagnostics to find outliers?

- Because they depend on pihat



What measures do we use to assess the predictive ability of a model?

- Compare each subject's predicted and observed outcomes and determine for how many study subjects the model "got it right"
- Do this by calculating Se, Sp, PPV and NPV for the model

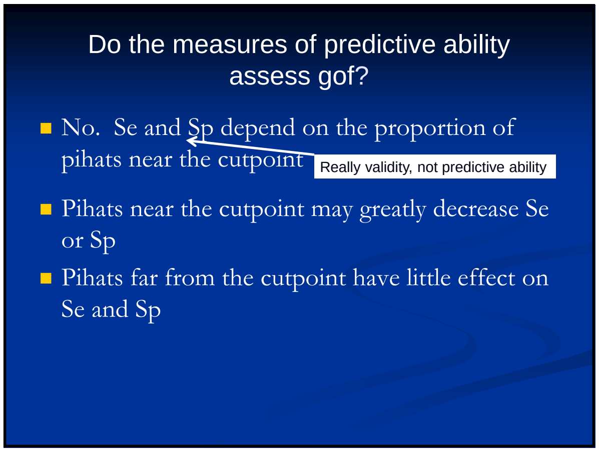
AND/OR

Really validity, not predictive ability

- Calculate the area under the ROC curve

Area under the ROC curve

- Pair each diseased subject with each non-diseased subject and compare each pair's pihats
- Determine the proportion of pairs where pihat for the diseased subject is greater than pihat for the non-diseased subject

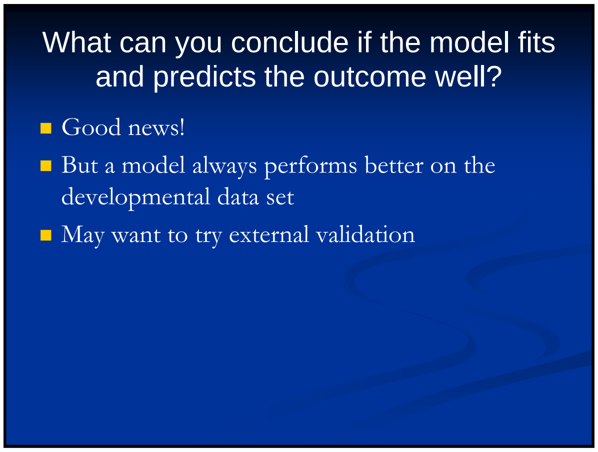


Do the measures of predictive ability assess gof?

- No. Se and Sp depend on the proportion of pihats near the cutpoint

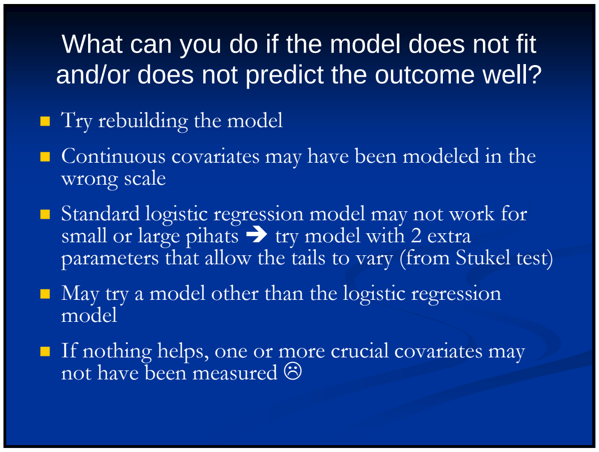
Really validity, not predictive ability

- Pihats near the cutpoint may greatly decrease Se or Sp
- Pihats far from the cutpoint have little effect on Se and Sp



What can you conclude if the model fits and predicts the outcome well?

- Good news!
- But a model always performs better on the developmental data set
- May want to try external validation



What can you do if the model does not fit and/or does not predict the outcome well?

- Try rebuilding the model
- Continuous covariates may have been modeled in the wrong scale
- Standard logistic regression model may not work for small or large pihats → try model with 2 extra parameters that allow the tails to vary (from Stukel test)
- May try a model other than the logistic regression model
- If nothing helps, one or more crucial covariates may not have been measured ☹