Final Study Guide

STA 363

For the final exam, you should be able to do the following:

1) Reduced F-tests

- Understand when they are used

    When we want to compare our full model to a "reduced" model, which has fewer predictors.

    Full model F-test is a special case where the reduced model is an intercept only model, notation in R: response ~ 1.

    Especially useful for testing multi-level factors (e.g. age categories), where we can test multiple dummy variables at once.

    Also useful for testing interactions.

    Note: ONLY FOR LINEAR MODELS. For GLM, use likelihood ratio test instead.

- Know what variables are being tested based on code

    Should be straightforward: whatever variables are left out of the reduced model.

    Note: we are NOT testing the variables present in both models.

- Understand how to interpret the output

    Look for the F-stat, degrees of freedom, and the p-value in the output.

2) Categorical predictors

- Understand coding of dummy variables
    - Binary predictors: *these are easier, just one dummy variable coded 0/1*

- Predictors with 3+ factor levels: *have to choose a "reference" category, and set up k-1 dummy variables, where k is the number of factor levels.*

  Interpretations of model coefficients are based on the reference category: all categories are compared to the reference.

  Since reference category is represented by all dummy variables equal to 0, intercept represents the group that is in the reference category for all categorical predictors.

- Understand how to interpret linear model coefficients for categorical predictors

  Main idea here: comparing groups. e.g. average difference in the response variable between males and females, or average difference between age 21-30 and age 11-20.

- ANCOVA:
  - Test for significant interactions

    We can use either the T-test in the model output or do a reduced F-test to test this.

    Leave interaction out of model if p-value is not less than 0.05 (prefer simpler model).

  - Understand when you can and cannot interpret main effects (like two-way ANOVA)

    Main effects are the coefficients for the non-interaction terms. (may need to review this term).

    If there is a significant interaction, cannot interpret main effects. Why? (Great question for the class to consider, also good review from two-way ANOVA)

  - Write fitted models for both levels of a binary predictor

    These equations are in the course notes and in the text.

Let X be the binary predictor and Z be the numerical predictor

*For X=0, fitted model is just beta_0 + beta_1*X*

*For X=1, fitted model is (beta_0 + beta_2) + (beta_1 + beta_3)*X*

(with interaction we get 2 intercept coefficients and two slope coefficients)

3) Model building

- Transformations
  - Box-Cox (understand how to interpret plot and what transformations are covered)

    Peak represents optimal power transformation (e.g. X^2 or sqrt(X))

    Zero actually corresponds to log transformation.

    If one is in the range of acceptable values, no transformation necessary

- Unusual Observations (high leverage points, outliers, influential points)
  - Be able to identify these

    What plot do we look at? (residuals vs. leverage)

    Outliers: residuals larger than +/- 3

    High-leverage: look for natural gaps in the leverage (x-axis) – can also compute a threshold, but not expected to memorize that formula

  - Know what can be done about them

    The main thing is to verify that they are legitimate data entries. If so, should not remove them.

    Can use a dummy variable to represent a single observation.

Can also fit model both ways, see if results are different.

- Multicollinearity
  - VIFs (>10 indicates a multicollinearity issue)

    Nothing more to say here

  - How to address (remove or scale predictors)

    Again, this one has the answer right there

    Scaling predictors means we standardize them by centering and scaling – every predictor is represented by Z-scores instead. Problem: harder to interpret.

4) Model Selection

- Model Selection criteria (be able to decide best model based on these)
  - AIC: *lower is better*
  - BIC: *lower is better*
  - R^2: *higher is better*
- Step-wise selection
  - Understand the starting models for forward and backward selection

    Forward: start with empty model. MUST ALSO SPECIFY SCOPE

    Backward: start with full model.

    By default, chooses models based on AIC

  - Interpret model output from stepwise selection output

Shows each iteration with AIC values as well as which variables were removed or added at each step

- Interpret model output from the chosen model

  Same as any other linear model output at this point.

  Review linear model output: F-test, T-tests, coefficients

- Best Subsets method
  - Understand how this is different from stepwise selection

    Checks every combination of predictors. Step-wise selection only checks some of the models

  - Main limitation for best subsets?

    Computation is slow

5) Cross-validation

- Benefits of model validation? (compared to other model selection criteria)

  Eliminates the bias that comes from using the same data for both fitting and for evaluation

- Understand the main concepts behind model validation

  Divide data into 2 parts:

  Training data: fit model (compute model coefficients)

  Test data: evaluate model (compute RMSE)

- Understand how cross-validation works

What does the number of folds control? *How many groups we create from the data for testing sets.*

- Choose models based on cross-validation output

  Check RMSE values

6) Logistic Regression

- Model form: *logit(p) = beta_0 + beta_1*X + …*

  Can also say "log odds" on the left side of this equation (logit is the function for log odds)

- Know the relationships between p, odds, and log odds

  Odds = p/(1-p) = P(Success)/P(Failure)

  Obviously log odds is just the log of this

- Interpret model coefficients (intercept and other coefficients)

  ***Exponentiated** intercept is the odds of [success] when all predictors are equal to 0. (This may include dummy variables, must know which factor level is the reference category.)*

  *Other coefficients (**when exponentiated as well**) represent odds ratios.*

  Must remember that effects are multiplicative. E.g. a two-unit increase in a predictor will increase the response by (e^beta)^2 times

- Use deviance to describe variability

  If the model is a good fit, null deviance should be large compared to residual deviance.

  Null deviance is basically total variation. Residual deviance is basically error variation. Want error to be relatively small in a "good" model.

7) Poisson Regression

- Model form: log(lambda) = *beta_0 + beta_1*X + …*
- Basic idea of Poisson Regression

8) Understand when to use any of the different models we have discussed over the semester

- ANOVA (One-way, Two-way, Blocked, Repeated Measures)
- Linear Regression (Simple and Multiple)
- Generalized Linear Models (Logistic Regression, Poisson Regression)