**Stat 500** **Lab 12** **Fall 2020**

Objectives:

(1) Analyze the heights (**heights.sas**) dataset (introducing categorical predictor variables in regression analysis).
(2) Analyze biomass dataset (**biomass.sas**) to illustrate SAS code for model selection.

**Difference in selection procedures:**

- Forward selection
    - Once a variable is added to model, variable stays in model.
    - When variables are highly correlated, this can result in non-significant variables in model.
- Backward elimination
    - All variables have a chance to stay in model.
    - Once a variable is eliminated from model, variable is not considered for re-entry.
- Stepwise selection
    - Variables must add significantly to model to stay in model.
- All Possible Models
    - Allows user to see some models not seen in variable selection procedures
    - Can be overwhelming to select model if large number of explanatory variables.

Multicollinearity

- Multicollinearity is the situation where one or more predictor variables are “nearly” linearly related to the others
- Problem can be a pair of highly correlated variables or a large group of moderately correlated variables
- It is not a violation of model assumptions

Effects of multicollinearity

- Fitted values are OK
- Estimated regression coefficients have high standard errors
- Difficult to interpret individual regression coefficients
- Regression coefficients change a lot with minor changes in model/data (e.g., if we remove a variable or case)

Remedies for multicollinearity

- Use the model only for prediction
- Drop variables that are highly correlated (need to be careful)
- Create composite (combined) variables (e.g., using principal components)
- Find new cases that “break” the observed correlation (i.e., have a different pattern)