Model Building

- **Overview of Model Building**. Four phases of model building:
  1. Data Collection
  2. Reduction of Explanatory/Predictor Variables
  3. Model Refinement / Model Selection
  4. Model Validation
- **Data Collection Phase.**
  - Studies can be *controlled experiments* or *observational studies*. It is important to know the difference, and know which of the two you are working with.
  - Studies have different types of variables, and it is again important to understand this prior to analysis.
- **Controlled Experiments.**
  - In a *controlled experiment*, levels of explanatory variables are controlled; a combination of these levels (called a *treatment*) is assigned to each experimental unit.
  - If completely controlled, one can get a *balanced design*, and hence no intercorrelation among predictors (the ideal situation).
  - Sometimes also have uncontrolled variables which are often called *covariates*
  - Usually a small number of variables involved.
  - Correct model is often 'obvious' and selection of explanatory variables unneeded.
  - May need to develop the correct functional form for the model (interactions, non- linear terms, etc.)
- **Observational Studies.**
  - In an observational study all variables are "observed" rather than "controlled" and hence multicollinearity is much more likely.
  - Study can be *exploratory* – we want to narrow a large group of explanatory variables to find those that are important.
  - Study could also be *confirmatory* – have specific hypotheses going in and collect exactly the data needed to test them.
  - Exploratory studies require the most in terms of model selection.
  - Start with a large number of potential explanatory variables (some will be intercorrelated) and reduce to a few appropriate subsets for final consideration.
- **Model Selection.**
  - How many variables to use? Smaller sets are more convenient (easier to interpret), but larger sets may explain more of the variation in the response.
  - Given a subset size, which variables should we choose for the model? Some statistics will help us compare them.

- Model Selection Criteria
  1. R^2/SSE
  2. Adjusted R^2/MSE
  3. Mallow's C_p Criterion
  4. AIC/SBC
  5. PRESS Statistics

- **R_p^2 (SSE_p) Criterion.**
  - The subscript p corresponds to the number of parameters in the model.

$$R_p^2 = 1 - \frac{SSE_p}{SSTO}$$

  - The goal is to maximize its value. However, as we already know, R_p^2 continues to go up as variables are added to the model. But eventually extra variables are just going to get in the way.
  - This criterion is useful in comparing models with same number of variables. Among these, the model with the highest R_p^2 could be considered "best".
  - Maximizing R_p^2 is equivalent to choosing the $(p-1)$-variable model with the smallest SSE_p.

- **Adjusted $R_{p,\text{adj}}^2$ (MSE) Criterion.**
  - Penalizes the R^2 value based on the number of variables in the model:

$$R_{p,\text{adj}}^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO} = 1 - \frac{MSE_p}{SSTO/(n-1)} = 1 - \frac{MSE_p}{\text{constant}}$$

  - End up subtracting off more if p gets larger, so if this is not counter balanced by enough decrease in SSE, $R_{p,\text{adj}}^2$ can decrease as a variables are added.
  - Maximizing the $R_{p,\text{adj}}^2$, or equivalently minimizing the MSE_p, is one way to determine a best model.
  - $R_{p,\text{adj}}^2$ has an advantage over regular R_p^2 since it compares models of different size.

- **Mallows' C_p Criterion.**
  - Basic idea: Compare predictive ability of subset models to that of full model
  - Full model generally best for prediction; but if multicollinearity is present then parameter estimates are not useful.
  - Subset of full model that doesn't have as much multicollinearity will be better as long as there is no substantial bias in predicted values relative to full model (i.e. close to same predictive ability.)
  - C_p considers ratio of the SSE for $p-1$ variable model to the MSE of the full model; then penalizes for the number of variables:

$$C_p = \frac{SSE_p}{MSE(full)} - (n - 2p)$$

- A model is considered 'good' if
  1. C_p is small.
  2. C_P is close to p.

- **AIC and SBC.**
  - AIC is Akaike's Information Criterion

$$AIC_p = n\ln\left(\frac{SSE_p}{n}\right) + 2p$$

  - SBC is Schwarz' Bayesian Criterion

$$SBC_p = n\ln\left(\frac{SSE_p}{n}\right) + p\ln(n)$$

  - We want to minimize these for "best model". They are the same except for their 'penalty' terms.

- **PRESS (Prediction Sum of Squares) Statistic.**
  - Measures how well fitted values predict the observed responses (SSE is a similar measure)
  - Different from SSE in that, for each observation, model based on data excluding that observation is used for the prediction

$$PRESS_p = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2$$

    $\hat{Y}_{i(i)}$ is the prediction for the ith observation using all data except ith observation.
  - Models with a small PRESS statistic are considered good candidate models.

- **Model Selection - Summary.**
  - Compare models of the same size using R^2 (maximize)
  - Compare different sized models using adjusted R^2 (max) or AIC/SBC (min)
  - Mallows' C_p Criterion for predictive ability of the model (minimize compared to p)
  - PRESS statistic for predictive ability (measures prediction error, smaller is better)
  - Smaller sets of variables are more convenient, particularly for parameter interpretation.
  - Larger sets of variables may explain more of the variation in the response and be better for prediction.

- **Stepwise Regression Methods.**
  1. *Forward Selection* - From group of variables that can be added, add to the model the one with the largest variable added-last t-statistic (or least p-value).
  2. *Backward Selection* - Start with full model and delete variables that "can" be deleted, one by one, starting with the smallest variable-added-last t-statistic (or largest p-value).
  3. *Stepwise* - Combine forward selection with backward elimination, checking for entry, then removal, until no more variables can be added or removed.

- **Model Validation.**
  - Mean Square Prediction Error:

$$MSPR = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

    1. Y_i = value of response in ith validation case.
    2. $\hat{Y}_i$ = =predicted value for ith validation case based on model selected in model-building data set.
    3. n=number of cases in the validation data set.
  - If MSPR is fairly close to MSE, this suggests the model is reasonable.
  - Or if the regression coefficients obtained using the validation data set is close to those obtained using the training data set, then the model is reasonable.

Model Building - Diagnostics

- **Outlier Detection in Multiple Linear Regression (MLR).**
  - We can have both X and Y outliers.
  - In SLR, outliers were relatively easy to detect via scatterplots or residual plots.
  - In MLR, it becomes more difficult to detect outlier via simple plots.
- **Detection of Y Outliers Using Residuals.**
  - We could use residuals to identify outlying values in Y (large magnitude implies extreme value)
  - But residuals don't really have a scale, so it's hard to know if it's large. We need something more standard.
- **Semi-studentized Residuals.**
  - Recall that $\epsilon_i \sim iidN(0, \sigma^2)$. Hence, ϵ_i/σ is standard normal.
  - However, we don't know the true errors or σ, so we use residuals and $\sqrt{MSE}$.
  - $e_i/\sqrt{MSE}$ is a *semi-studentized residual.*
- **Studentized Residuals.**
  - The actual standard deviation of the residual is $s\{e\} = \sqrt{MSE(1-h_{ii})}$.
  - Where h_{ii} is the ith element on the main diagonal of the hat matrix, $H = X(X'X)^{-1}X'$. This value is between 0 and 1.
  - $e_i^* = \dfrac{e_i}{\sqrt{MSE(1-h_{ii})}} \sim t_{(n-p)}$ is a *studentized residual.*
- **Studentized Deleted Residuals (SDR).**
  - Each residual is obtained by regressing using all of data except for the point in question.
  - Similar to what is done to compute PRESS statistic: $d_i = Y_i - \hat{Y}_{i(i)}$
  - To avoid computing entire regression over and over again, one can use the formula:

$$d_i = \frac{e_i}{(1-h_{ii})}$$

  - The standard deviation for this residual is

$$s\{d_i\} = \sqrt{\frac{MSE_{(i)}}{1-h_{ii}}}$$

  - $t_i = \dfrac{d_i}{s\{d_i\}} = \dfrac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \sim t_{(n-p-1)}$ is called the *studentized deleted residual.*
  - Alternative formula to calculate these without rerunning the regression n times

$$t_i = e_i\sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}}$$

  - Many high SDR indicates inadequate model.

- **Leverage Values.**
  - Outliers in X can be identified because they will have large leverage values. The leverage is just h_{ii} from the hat matrix.
  - Note: $0 \leq h_{ii} \leq 1$ and $\sum h_{ii} = p$.
  - Large leverage values indicate the ith case is distant from the center of all X obs.
  - Leverage considered large if it is bigger than twice the mean leverage value, $2p/n$.
- **Other Influence Statistics.**
  - Not all outliers have a strong influence on the fitted model. Some measures to detect the influence of each observation are:
  - *Cook's Distance* measures the influence of an observation on all fitted values. Large values suggest that an observation has a lot of influence (can compare to an $F(p, n-p)$ distribution).
  - *DFFits* measures the influence of an observation on its own fitted value.
    - Essentially measures difference between prediction of itself with/without using that observation in the computation.
    - Large absolute values (bigger than 1, or bigger than $2\sqrt{p/n}$) suggest that an observation has a lot of influence on its own prediction.
  - *DFBeta* measures the influence of an observation on a particular regression coefficient. Absolute values bigger than 1 or $2/\sqrt{n}$ are considered large.