Regression Model Building

- Setting: Possibly a large set of predictor variables (including interactions).
- Goal: Fit a parsimonious model that explains variation in Y with a small set of predictors
- Automated Procedures and all possible regressions:
  - Backward Elimination (Top down approach)
  - Forward Selection (Bottom up approach)
  - Stepwise Regression (Combines Forward/Backward)
  - C_p Statistic - Summarizes each possible model, where "best" model can be selected based on statistic

Backward Elimination

- Select a significance level to stay in the model (e.g. SLS=0.20, generally .05 is too low, causing too many variables to be removed)
- Fit the full model with all possible predictors
- Consider the predictor with lowest *t*-statistic (highest *P*-value).
  - If P > SLS, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change)
  - If $P \leq$ SLS, stop and keep current model
- Continue until all predictors have *P*-values below SLS

Forward Selection

- Choose a significance level to enter the model (e.g. SLE=0.20, generally .05 is too low, causing too few variables to be entered)
- Fit all simple regression models.
- Consider the predictor with the highest *t*-statistic (lowest *P*-value)
  - If $P \leq$ SLE, keep this variable and fit all two variable models that include this predictor
  - If $P >$ SLE, stop and keep previous model
- Continue until no new predictors have $P \leq$ SLE

Stepwise Regression

- Select SLS and SLE (SLE<SLS)
- Starts like Forward Selection (Bottom up process)
- New variables must have $P \leq$ SLE to enter
- Re-tests all “old variables” that have already been entered, must have $P \leq$ SLS to stay in model
- Continues until no new variables can be entered and no old variables need to be removed

All Possible Regressions - C_p

- Fits every possible model. If K potential predictor variables, there are 2^K-1 models.
- Label the Mean Square Error for the model containing all K predictors as MSE_K
- For each model, compute SSE and C_p where p is the number of parameters (including intercept) in model

$$C_p = \frac{SSE}{MSE_K} - (n - 2p)$$

- Select the model with the fewest predictors that has $C_p \approx p$

Regression Diagnostics

- Model Assumptions:
  - Regression function correctly specified (e.g. linear)
  - Conditional distribution of *Y* is normal distribution
  - Conditional distribution of *Y* has constant standard deviation
  - Observations on *Y* are statistically independent
- Residual plots can be used to check the assumptions
  - Histogram (stem-and-leaf plot) should be mound-shaped (normal)
  - Plot of Residuals versus each predictor should be random cloud
    - U-shaped (or inverted U) $\Rightarrow$ Nonlinear relation
    - Funnel shaped $\Rightarrow$ Non-constant Variance
  - Plot of Residuals versus Time order (Time series data) should be random cloud. If pattern appears, not independent.

Detecting Influential Observations

- **Studentized Residuals** – Residuals divided by their estimated standard errors (like *t*-statistics). Observations with values larger than 3 in absolute value are considered outliers.
- **Leverage Values (Hat Diag)** – Measure of how far an observation is from the others in terms of the levels of the independent variables (not the dependent variable). Observations with values larger than 2(k+1)/n are considered to be potentially highly influential, where k is the number of predictors and n is the sample size.
- **DFFITS** – Measure of how much an observation has effected its fitted value from the regression model. Values larger than 2*sqrt((k+1)/n) in absolute value are considered highly influential. Use standardized DFFITS in SPSS.

Detecting Influential Observations

- **DFBETAS** – Measure of how much an observation has effected the estimate of a regression coefficient (there is one DFBETA for each regression coefficient, including the intercept). Values larger than 2/sqrt(n) in absolute value are considered highly influential.
- **Cook's D** – Measure of aggregate impact of each observation on the group of regression coefficients, as well as the group of fitted values. Values larger than 4/n are considered highly influential.
- **COVRATIO** – Measure of the impact of each observation on the variances (and standard errors) of the regression coefficients and their covariances. Values outside the interval 1 +/- 3(k+1)/n are considered highly influential.

Obtaining Influence Statistics and Studentized Residuals in SPSS

– .Choose **ANALYZE**, **REGRESSION**, **LINEAR**, and input the Dependent variable and set of Independent variables from your model of interest (possibly having been chosen via an automated model selection method).
  - **.**Under **STATISTICS**, select **Collinearity Diagnostics**, **Casewise Diagnostics** and **All Cases** and **CONTINUE**
  - **.**Under **PLOTS**, select **Y:*SRESID** and **X:*ZPRED**. Also choose **HISTOGRAM**. These give a plot of studentized residuals versus standardized predicted values, and a histogram of standardized residuals (residual/sqrt(MSE)). Select **CONTINUE**.
  - **.**Under **SAVE**, select **Studentized Residuals**, **Cook's**, **Leverage Values**, **Covariance Ratio**, **Standardized DFBETAS**, **Standardized DFFITS**. Select **CONTINUE**. The results will be added to your original data worksheet.

Variance Inflation Factors

- **Variance Inflation Factor (VIF)** – Measure of how highly correlated each **independent variable** is with the other predictors in the model. Used to identify **Multicollinearity**.
- Values larger than 10 for a predictor imply large inflation of standard errors of regression coefficients due to this variable being in model.
- Inflated standard errors lead to small *t*-statistics for partial regression coefficients and wider confidence intervals

Nonlinearity: Polynomial Regression

- When relation between Y and X is not linear, polynomial models can be fit that approximate the relationship within a particular range of X
- General form of model:

$$E(Y) = \alpha + \beta_1 X + \cdots + \beta_k X^k$$

- Second order model (most widely used case, allows one "bend"):

$$E(Y) = \alpha + \beta_1 X + \beta_2 X^2$$

- Must be very careful not to extrapolate beyond observed X levels

Generalized Linear Models (GLM)

- General class of linear models that are made up of 3 components: Random, Systematic, and Link Function
  - Random component: Identifies dependent variable (*Y*) and its probability distribution
  - Systematic Component: Identifies the set of explanatory variables ($X_1,\ldots,X_k$)
  - Link Function: Identifies a function of the mean that is a linear function of the explanatory variables $g(\mu) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$

Random Component

- Conditionally *Normally* distributed response with constant standard deviation - Regression models we have fit so far.
- Binary outcomes (Success or Failure)- Random component has *Binomial* distribution and model is called **Logistic Regression**.
- Count data (number of events in fixed area and/or length of time)- Random component has *Poisson* distribution and model is called **Poisson Regression**
- Continuous data with skewed distribution and variation that increases with the mean can be modeled with a *Gamma* distribution

Common Link Functions

- Identity link (form used in *normal* and *gamma* regression models): $g(\mu) = \mu$

- Log link (used when μ cannot be negative as when data are *Poisson* counts): $g(\mu) = \log(\mu)$

- Logit link (used when μ is bounded between 0 and 1 as when data are binary):
$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

Exponential Regression Models

- Often when modeling growth of a population, the relationship between population and time is exponential: $E(Y) = \mu = \alpha\beta^X$

- Taking the logarithm of each side leads to the linear relation: $\log(\mu) = \log(\alpha) + X\log(\beta) = \alpha' + \beta' X$
- Procedure: Fit simple regression, relating log(*Y*) to *X*. Then transform back:

$$\widehat{\log(Y)} = a + bX \qquad \hat{\alpha} = e^a \qquad \hat{\beta} = e^b \qquad \hat{Y} = \hat{\alpha}\hat{\beta}^X$$