STAT 500

Model Diagnostics for
Multiple Linear Regression Models

MLR Model and Assumptions

$$Y_i = \mu_{Y|\mathbf{x}} + \epsilon_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \text{ where } \epsilon_i \text{ i.i.d. } N(0, \sigma^2)$$

- Observations Y_i are independent.
- Values of $\mathbf{x}$ are fixed.
- $\mu_{Y|\mathbf{x}}$ is a linear function of $\mathbf{x}$
- Homogeneous error variance: $Var(\epsilon_i) = \sigma^2$
- Normally distributed errors: ϵ_i i.i.d. $N(0, \sigma^2)$

Independence

- Check independence of observations through details of data collection.
- Beware of
  - Observations over time
  - Clustering of observations
  - Spatial elements to observations
- Crucial assumption - must use other methods if violated.

Fixed Values of $\mathbf{x}$

- Assume $\mathbf{x}$ is measured without error.

- Check through definition of variables and through details of data collection.

- If violated for some x_j, model the error in those x_j using random effects.

Linearity

- Scatterplot - Plot of Y versus each x_j
  - Linear Patterns

- Residual Plot - Plot of residuals $\mathbf{e}$ versus each x_j
  - No patterns

Violations of Linearity

- Transform Y values so that relationship with each x_j is linear.
- Transform each of the x_j variables to have linear relationship with Y.
- Common transformations:
  - Power: Y^2, Y^3, $\sqrt{Y}$, etc.
  - Exponential: $\exp(Y)$, $\ln(Y)$
- Conduct analysis with transformed Y and/or $\mathbf{x}$ values.
- Undo transformation in drawing conclusions.

Homogeneous Variance

- Residual Plots - scatterplots of residuals with predicted values $\hat{Y}$ and with each x_j
  - Look for changes in variability around the horizontal line at 0.
  - Megaphone shaped pattern: variability of e increases or decreases as either $\hat{Y}$ or specific x_j increases.
- Impact: Confidence Intervals for Conditional Mean and Prediction Intervals

Violations of Homogeneous Variance

- Transform Y or x_j
- Use Weighted Least Squares

Weighted Least Squares

- Assume $Var(\epsilon_i) = \sigma_i^2$ for $i = 1, \ldots, n$
- Define diagonal matrix W to have elements $w_{ii} = 1/\sigma_i^2$
- Weighted least squares estimate of $\boldsymbol{\beta}$ is

$$(X^T W X)^{-1} X^T W Y$$

Weighted Least Squares

- Observations with smaller σ_i^2 get a larger weight in the weighted least squares estimate than observations with larger σ_i^2.

- Must know or be able to estimate values of w_{ii}.

  - If the ith observation is an average of n_i equally variable observations, then $Var(Y_i) = \sigma^2/n_i$ and $w_{ii} = n_i$.

  - If the ith observation is a total of n_i observations, then $Var(Y_i) = n_i\sigma^2$ and $w_{ii} = 1/n_i$.

  - If variance is proportional to some predictor x_j, then $Var(Y_i) = x_{ij}\sigma^2$ and $w_{ii} = 1/x_{ij}$.]

  - In some cases, the values of the weights may be based on theory or prior research.

Weighted Least Squares

- The difficulty in applying weighted least squares in practice is determining the weights (estimate of error variances).

- Estimation schemes exist for estimating weights based on other characteristics (megaphone shape or upward trend in residual plots)

- Least squares and weighted least squares estimates are usually similar in value.

- Differences occur with inference and prediction.

Normality

- Distribution of Residuals
  - Histogram of residuals
  - Normal probability plot of residuals
  - Tests for normality of residuals

- Affects inference, especially for smaller sample sizes

Violations of Normality

- Remedies
  - Check for outliers
  - Transform Y
  - Conduct robust regression

Model Selection Assessment

- Multiple Testing Problem
  - Adjust significance for all models considered?
  - Ignore the issue (most common practice) i.e. assume selected model is correct
  - Explore conclusions from several of the best models
  - Model averaging (using AIC or BIC)

Model Selection Assessment

- Stepwise procedures tend to overfit the sample data. Would the model perform as well in making predictions for new cases randomly selected from the population?
- Model validation: Split data into two parts
  - Training sample (perhaps 2/3 of the data)
  - Validation sample (the remainder of the data)
  - Use training sample to select model
  - Use validation sample to assess model performance and fit

Model Selection Assessment

- Compute

$$MSE_{\text{Validation}} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (Y_i - \widehat{Y}_i)^2$$

- Should be approximately equal to MSE_{Training} from selected model
- $MSE_{\text{Validation}}$ will be substantially larger if model is over fit to the training sample
- Use as model selection technique - fit many models to the training sample and compute $MSE_{\text{Validation}}$ for the validation sample for each selected model
- PRESS is doing this over-and-over with the size of the test sample equal to one case

Model Selection: PRESS Criterion

- PRESS (predicted residual error sum of squares)
  - Predict each response using the other $n-1$ cases to estimate the model parameters
  - $PRESS = \sum_i (Y_i - \widehat{Y}_{i(-i)})^2$
  - nice idea, not used very often

Case Diagnostics

- Leverage
- Outliers
- Influential Points

Case Diagnostics - Leverage

- Extreme values in $\mathbf{x}$'s are called high leverage cases because they exert a large "pull" on the fitted regression model
- Measured using the projection matrix P_X (also called the hat matrix $= H$)

$$\hat{\mathbf{Y}} = H\mathbf{Y} = P_X\mathbf{Y} = X(X^TX)^{-1}X^TY$$

Case Diagnostics - Leverage

- For an observation i, can write

$$\widehat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j$$

- h_{ii} is the (i, i) element of P_X and it is called the *leverage* of the $i-$th case.

- The leverage measures the extent to which ith observation dictates its own fitted value

Case Diagnostics - Leverage

- Properties of h_{ii}
  - $0 \leq h_{ii} \leq 1$
  - $\sum_{i=1}^{n} h_{ii} = k + 1$
  - Measures the "distance" between the vector of values for the explanatory variables for the ith observations and the average vector of values of explanatory variables

Case Diagnostics - Leverage

- Often use $2(k+1)/n$ or $3(k+1)/n$ as a guide for determining large h_{ii}

- In addition to an absolute cutoff, look for large h_{ii} by examining the distribution of h_{ii} values across cases

Case Diagnostics - Outliers

- Extreme Y_i value for a given $\mathbf{x}$
- Three assessment methods
  - Residuals
  - Internally studentized residuals
  - Externally studentized residuals

Case Diagnostics - Residuals

- Residuals

$$e_i = Y_i - \widehat{Y}_i$$

- $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$
- Observations with higher leverage will have residuals with smaller variance.

Case Diagnostics - Residuals

- Internally studentized residuals

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

- r_i will have mean zero and approximately equal variance

- Outliers will inflate MSE

- r_i is called STUDENT in SAS

Case Diagnostics - Residuals

- Externally studentized residuals

$$t_i = \frac{e_i}{\sqrt{MSE_{(-i)}(1 - h_{ii})}}$$

  where $\text{MSE}_{(-i)}$ is MSE without the ith observation

- t_i will have mean zero and approximately equal variance

- t_i is called RSTUDENT in SAS

Case Diagnostics - Outliers

Residual values with absolute value

- Less than 2 are fine
- Between 2 and 3 indicate potential outliers
- Greater than 3 indicate outliers

Case Diagnostics - Outliers

- Outliers inflate value of $\hat{\sigma}^2$

- Will lower values of t and F test statistics

- Will inflate widths of confidence intervals for parameters and prediction intervals

Case Diagnostics - Influence

- Concerned about unusual cases that have a big influence on both:
  - $\widehat{Y}_i$ for some $\mathbf{x}_i$
  - regression coefficient $\widehat{\beta}_j$
- Could delete the case, refit model and examine the change

Case Diagnostics - Influence

- COOK'S D - effect deleting the i-th case on the entire set of fitted values

$$D_i = \frac{\sum_j (\widehat{Y}_j - \widehat{Y}_{j(-i)})^2}{(k+1)MSE} = \left(\frac{r_i^2}{k+1}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

- D_i is large when r_i is large and h_{ii} is large

- There is no gold-standard for the cutoff of Cook's D.
  - SAS uses $4/n$.
  - $D_i > 2 * \sqrt{2/n}$ indicates substantial influence.
  - $D_i > F_{k+1,n-k-1,0.5}$ indicates substantial influence.
  - Can also judge D_i relative to other D_j's.

Case Diagnostics - Influence

- DFFITS_i - effect of ith case on fitted value for Y_i

$$\text{DFFITS}_i = \frac{\widehat{Y}_i - \widehat{Y}_{i(-i)}}{\sqrt{MSE_{(-i)} h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

- $|\text{DFFITS}_i| > 2$ is considered large in small or medium sized samples
- $|\text{DFFITS}_i| > 2\sqrt{\frac{k+1}{n}}$ is considered large in big samples

Case Diagnostics - Influence

- DFBETAS - effect of deleting the i-th case on the estimate of a single coefficient

$$\text{DFBETA}_{k,i} = \frac{b_k - b_{k(-i)}}{\sqrt{MSE_{(-i)} c_{kk}}}$$

- $k = 0$ for population intercept β_0

- $k = j$ for population slope β_j

- c_{kk} is (k, k) element of $(X^T X)^{-1}$

- DFBETA larger than 2 (small or medium size samples) or larger than $2n^{-1/2}$ (large samples) may be worthy of attention

Case Diagnostics - SAS

```
/* case diagnostics*/
 proc reg data=set1 plots=(diagnostics);
    model y = x1-x5/ vif influence;
    run;
```