Model Building III – Remedial Measures

KNNL – Chapter 11

Unequal Error Variances – Weighted Least Squares (WLS)

- Case 1 – Error Variances known exactly (*very* rare case)
- Case 2 – Error Variances known up to a constant
  - Occasionally information known regarding experimental units regarding the relative magnitude (unusual)
  - If “observations” are means of different numbers of units (each with equal variance) at the various X levels, Variance of observation *i* is σ^2/n_i where n_i is known
- Case 3 – Variance (or Standard Deviation) is related to one or more predictors, and relation can be modeled (see Breusch-Pagan Test in Chapter 3)
- Case 4 – Ordinary Least Squares with correct variances

WLS – Case 1 Known Variances - I

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad \varepsilon_i \sim N\left(0, \sigma_i^2\right) \quad i = 1, \ldots, n \qquad \sigma\left\{\varepsilon_i, \varepsilon_j\right\} = 0 \quad \forall\, i \neq j$$

$$\Rightarrow \boldsymbol{\sigma}^{\mathbf{2}}\left\{\varepsilon\right\} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Maximum Likelihood Estimation:

$$L\left(\beta\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2}\left(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1}\right)^2\right] \text{ setting: } w_i = \frac{1}{\sigma_i^2}$$

$$\Rightarrow \quad L\left(\beta\right) = \left[\prod_{i=1}^{n} \sqrt{\frac{w_i}{2\pi}}\right] \exp\left[-\frac{1}{2}\sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1}\right)^2\right]$$

To maximize $L\left(\beta\right)$, we need to minimize $Q_w = \sum_{i=1}^{n} w_i \left(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1}\right)^2$

Note that values with smaller σ_i^2 have larger weights w_i in the weighted least squares criterion.

WLS – Case 1 Known Variances - II

Easiest to set up in matrix form, where: $\mathbf{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix} = \mathbf{W}'$

$\sigma^{2}\left\{\mathbf{Y}\right\} = \sigma^{2}\left\{\varepsilon\right\} = \mathbf{W}^{-1}$

Normal Equations: $\left(\mathbf{X'WX}\right)\mathbf{b}_w = \mathbf{X'WY}$

$$\Rightarrow \quad \mathbf{b}_w = \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'WY} = \mathbf{AY} \quad \mathbf{A} = \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'W}$$

$$\Rightarrow \quad \mathbf{E}\left\{\mathbf{b}_{\mathbf{w}}\right\} = \mathbf{AE}\left\{\mathbf{Y}\right\} = \mathbf{AX}\beta = \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'WX}\beta = \beta$$

$$\Rightarrow \quad \sigma^{2}\left\{\mathbf{b}_{\mathbf{w}}\right\} = \mathbf{A}\sigma^{2}\left\{\mathbf{Y}\right\}\mathbf{A'} = \left(\mathbf{X'WX}\right)^{-1}\mathbf{X'WW}^{-1}\mathbf{WX}\left(\mathbf{X'WX}\right)^{-1} = \left(\mathbf{X'WX}\right)^{-1}$$

WLS – Case 2 – Variance Known up to Constant - I

Data are means of unequal numbers of replicates at X-levels, can use any weights generally

$$\sigma^2\left\{\overline{Y}_i\right\}=\sigma_i^2=\frac{\sigma^2}{r_i} \quad \Rightarrow \quad \boldsymbol{\sigma}^{\mathbf{2}}\left\{\mathbf{Y}\right\}=\sigma^2\begin{bmatrix} r_1^{-1} & 0 & \cdots & 0 \\ 0 & r_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_n^{-1}\end{bmatrix} \qquad r_i \equiv \text{ number of replicates at } i^{th} \text{ level of } X$$

$$w_i=\frac{1}{\sigma_i^2}=\frac{n_i}{\sigma^2} \quad \Rightarrow \quad \mathbf{W}=\frac{1}{\sigma^2}\begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_n\end{bmatrix}=\frac{1}{\sigma^2}\mathbf{W}^* \qquad \mathbf{W}^*=\begin{bmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_n\end{bmatrix}$$

$$\Rightarrow \mathbf{X'WX}=\mathbf{X'}\frac{1}{\sigma^2}\mathbf{W}^*\mathbf{X}=\frac{1}{\sigma^2}\mathbf{X'W}^*\mathbf{X} \Rightarrow \left(\mathbf{X'WX}\right)^{\mathbf{-1}}=\sigma^2\left(\mathbf{X'W}^*\mathbf{X}\right)^{\mathbf{-1}}$$

$$\Rightarrow \mathbf{X'WY}=\mathbf{X'}\frac{1}{\sigma^2}\mathbf{W}^*\mathbf{Y}=\frac{1}{\sigma^2}\mathbf{X'W}^*\mathbf{Y} \quad \Rightarrow \mathbf{b}_w=\left(\mathbf{X'WX}\right)^{\mathbf{-1}}\mathbf{X'WY}=\left(\mathbf{X'W}^*\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X'W}^*\mathbf{Y}$$

$$\boldsymbol{\sigma}^{\mathbf{2}}\left\{\mathbf{b}_w\right\}=\left(\mathbf{X'WX}\right)^{\mathbf{-1}}=\sigma^2\left(\mathbf{X'W}^*\mathbf{X}\right)^{\mathbf{-1}} \qquad \mathbf{s}^{\mathbf{2}}\left\{\mathbf{b}_w\right\}=MSE_w\left(\mathbf{X'W}^*\mathbf{X}\right)^{\mathbf{-1}} \qquad MSE_w=\frac{\sum_{i=1}^{n} w_i^*\left(Y_i-\hat{Y}_i\right)^2}{\left(n-p\right)}$$

WLS - Case 3 – Estimated Variances

- Use squared residuals (estimated variances) or absolute residuals (estimated standard deviations) from OLS to model their levels as functions of predictor variables
  - Plot Residuals, squared residuals and absolute residuals versus fitted values and predictors
  - Using model building techniques used on *Y* previously to obtain a model for either the variances or standard deviations as functions of the *X*'s or the mean
  - Fit estimated WLS with the estimated weights (1/variances)
  - Iterate until regression coefficients are stable (iteratively re-weighted least squares)

$$\hat{w}_i = \frac{1}{\hat{v}_i} = \frac{1}{\left(\hat{s}_i\right)^2} \qquad \mathbf{b}_{\hat{w}} = \left(\mathbf{X'}\hat{\mathbf{W}}\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X'}\hat{\mathbf{W}}\mathbf{Y} \qquad MSE_{\hat{w}} = \frac{\left(\mathbf{Y}-\mathbf{X}\mathbf{b}_{\hat{w}}\right)'\hat{\mathbf{W}}\left(\mathbf{Y}-\mathbf{X}\mathbf{b}_{\hat{w}}\right)}{n-p}$$

Case 4 – OLS with Estimated Variances

$$\mathbf{E}\left\{\mathbf{Y}\right\} = \mathbf{X\beta} \qquad \boldsymbol{\sigma}^{2}\left\{\mathbf{Y}\right\} = \boldsymbol{\sigma}^{2}\left\{\boldsymbol{\varepsilon}\right\} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\mathbf{b} = \left(\mathbf{X'X}\right)^{\mathbf{-1}} \mathbf{X'Y} = \mathbf{AY} \qquad \mathbf{A} = \left(\mathbf{X'X}\right)^{\mathbf{-1}} \mathbf{X'}$$

$$\Rightarrow \quad \boldsymbol{\sigma}^{2}\left\{\mathbf{b}\right\} = \mathbf{A}\boldsymbol{\sigma}^{2}\left\{\mathbf{Y}\right\}\mathbf{A'} = \left(\mathbf{X'X}\right)^{\mathbf{-1}} \mathbf{X'}\boldsymbol{\sigma}^{2}\left\{\boldsymbol{\varepsilon}\right\}\mathbf{X}\left(\mathbf{X'X}\right)^{\mathbf{-1}}$$

Note: $E\left\{\varepsilon_i^2\right\} = \sigma_i^2$ Use e_i^2 as an estimator of σ_i^2 : $\mathbf{S_0} = \begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{bmatrix}$

$$\mathbf{s}^{2}\left\{\mathbf{b}\right\} = \left(\mathbf{X'X}\right)^{\mathbf{-1}} \mathbf{X'S_0X}\left(\mathbf{X'X}\right)^{\mathbf{-1}}$$

This is referred to White’s estimator, and is robust to specification of form of error variance

Multicollinearity – Remedial Measures - I

- Goal: Prediction – Multicollinearity not an issue if new cases have similar multicollinearity among predictors
  - Firm uses shipment size (number of pallets) and weight (tons) to predict unloading times. Future shipments have similar correlation between predictors, predictions and PI^s are valid
- Linear Combinations of Predictors can be generated that are uncorrelated (Principal Components)
  - Good: Multicollinearity gone
  - Bad: New variables may not have “physical” interpretation
- Use of Cross-Section and Time Series in Economics
  - Model: Demand = f(Income , Price)
  - Step 1: Cross-section (1 time): Regress Demand on Income (b_I)
  - Step 2: Time Series: Regress Residuals from 1 on Price (b_P)

Multicollinearity – Ridge Regression -I

- Mean Square Error of an Estimator = Variance + Bias^2
- Goal: Add Bias to estimator to reduce MSE(Estimator)
- Makes use of Standard Regression Model with Correlation Transformation

$$MSE\left\{b^R\right\} = E\left\{\left(b^R - \beta\right)^2\right\} = \sigma^2\left\{b^R\right\} + \left(E\left\{b^R\right\} - \beta\right)^2 = \text{Variance + (Bias)}^2$$

Correlation Transformation / Standardized Regression Model:

$$Y_i^* = \frac{Y_i - \overline{Y}}{\left(\sqrt{n-1}\right)s_Y} \qquad X_{ik}^* = \frac{X_{ik} - \overline{X}_k}{\left(\sqrt{n-1}\right)s_k} \qquad Y_i^* = \beta_1^* X_{i1}^* + \ldots + \beta_{p-1}^* X_{i,p-1}^* + \varepsilon_i^* \quad \Rightarrow \quad \mathbf{r}_{XX}\mathbf{b} = \mathbf{r}_{YX}$$

Ridge Regression Estimator (with $c \geq 0$): $\left(\mathbf{r}_{XX} + c\mathbf{I}\right)\mathbf{b}^R = \mathbf{r}_{YX} \quad \Rightarrow \quad \mathbf{b}^R = \left(\mathbf{r}_{XX} + c\mathbf{I}\right)^{\mathbf{-1}}\mathbf{r}_{YX} = \begin{bmatrix} b_1^R \\ \vdots \\ b_{p-1}^R \end{bmatrix}$

Goal: choose small *c* such that the estimators stabilize (flatten) and VIF^s get small

Multicollinearity – Ridge Regression - II

Computational Formulas:

Variance Inflation Factors:

Diagonal Elements of the matrix:

$$\left(\mathbf{r}_{XX}+c\mathbf{I}\right)^{-1}\mathbf{r}_{XX}\left(\mathbf{r}_{XX}+c\mathbf{I}\right)^{-1}$$

Error Sum of Squares (Correlation Transformed Data):

$$\hat{Y}_i^* = b_1^R X_{i1}^* + \ldots + b_{p-1}^R X_{i,p-1}^* \qquad SSE_R = \sum_{i=1}^{n}\left(Y_i^* - \hat{Y}_i^*\right)^2$$

$$R_R^2 = 1 - SSE_R \qquad \text{since } \; SSTO_R = 1$$

Robust Regression for Influential Cases

- Influential cases, after having been ruled out as recording errors or indicative of new predictors, can be reduced in terms of their impacts on the regression model
- Two commonly applied Robust Regression Models:
  - Least Absolute Residuals (LAR) Regression – Choose the coefficients that minimize sum of absolute deviations (as opposed to squared deviations in OLS). No closed form solutions, must use specialized programs (can be run in R)
  - IRLS Robust Regression – Uses Iteratively Re-weighted Least Squares where weights are based on how much each case is an outlier (lower weights for larger outliers)

IRLS – Robust Regression

1. Choose weight function (we will use Huber here)
2. Obtain starting weights for each case.
3. Use starting weights in Weighted Least Squares and obtain residuals after fitting equation
4. Use Residuals from step 3 to generate new weights.
5. Continue iterations to convergence (small changes in estimated regression coefficients or residuals or fitted values)

Applying Procedure for Huber wight function (1.345 is tuning parameter to achieve high efficiency to normal model)

1. Huber weight function: $w = \begin{cases} 1 & |u| \leq 1.345 \\ \dfrac{1.345}{|u|} & |u| \leq 1.345 \end{cases}$

2. Use initial residuals from OLS fit after scaling by Median Absolute Deviation (Robust estimate of σ):

$$MAD = \frac{1}{0.6745}\text{median}\left\{\left|e_i - \text{median}\left\{e_i\right\}\right|\right\} \qquad u_i = \frac{e_i}{MAD}$$

3. Iterate to Convergence

Note: Bisquare weight Function: $w = \begin{cases} \left[1 - \left(\dfrac{u}{4.685}\right)^2\right]^2 & |u| \leq 4.685 \\ 0 & |u| \leq 4.685 \end{cases}$

Nonparametric Regression

- Locally Weighted Regressions (Lowess)
  - Works best with few predictors, and when data have been transformed to normality with constant error variances, and predictors have been scaled by their standard deviation (sqrt(MSE) or MAD when outliers are present)
  - Applies WLS with weights as distances from the individual points in a neighborhood to the target (q = proportion of data used in the neighborhood, typically 0.40 to 0.60)
- Regression Trees
  - X-space is broken down into multiple sub-spaces (1-step at a time), and each region is modeled by the mean response
  - Each step is chosen to minimize the within region sum of squares

Lowess Method

- Assumes model has been selected and built so assumptions of errors being approximately normal with constant variance holds.
- Predictors of different units should be scaled by SD or MAD in distance measure

Assuming p-1=2 predictor variables and

$q \equiv$ proportion of sample used at each point:

Distance Measure (Each point from target point): $d_i = \sqrt{\left(\frac{X_{i1}-X_{h1}}{s_1}\right)^2 + \left(\frac{X_{i2}-X_{h2}}{s_2}\right)^2}$

Weight Function: $w_i = \begin{cases} \left[1-\left(d_i/d_q\right)^3\right]^3 & d_i < d_q \\ 0 & d_i \geq d_q \end{cases}$

$d_q \equiv$ size of neighborhood so q of sample is included

Local Fitting: Fit first-order or second-order model by WLS and get fitted value at each target point

Regression Trees

- Splits region covered by the predictor variables into increasing number of sub-regions, where predicted value is mean response of all cases falling in the sub-region.
- Goal: Choose “cut-points” that minimize Error Sum of Squares for the current number of sub-regions – computationally intensive, no closed form solution
- Problem: SSE will continually drop until there are as many sub-regions as distinct combinations of predictors (SSE → 0)
- Validation samples can be used to determine which number of nodes (sub-regions – 1) that minimizes Mean Square Prediction Error (MSPR)

Bootstrapping to Estimate Precision

- Computationally intensive methods used to estimate precision of estimators in non-standard situations
- Based on re-sampling (with replacement) from observed samples, re-computing estimates repeatedly (nonparametric). Parametric methods also exist (not covered here)
- Once original sample is observed, and quantity of interest estimated, many samples of size n (selected with replacement) are obtained, each sample providing a new estimate.
- The standard deviation of the new sample quantities represents an estimate of the standard error

Two Basic Approaches

- Fixed X Sampling (Model is good fit, constant variance, predictor variables are fixed, i.e. controlled experiment)
  - Fit the regression, obtain all fitted values and residuals
  - Keeping the corresponding X-level(s) and fitted values, re-sample the n residuals (with replacement)
  - Add the bootstrapped residuals to the fitted values, and re-fit the regression (repeat process many times)
- Random X Sampling (Not sure of adequacy of model: fit, variance, random predictor variables)
  - After fitting regression, and estimating quantities of interest, sample n cases (with replacement) and re-estimate quantities of interest with “new” datasets (repeat many times)

Bootstrap Confidence Intervals – Reflection Method

Suppose we are interested in Estimating β_1 in simple regression (generalizes to other parameters):

1) Fit regression on original dataset, obtain $b_1 = \dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$

2) Obtain m bootstrap samples and their estimated slopes: $b_{j1}^* \quad j = 1,\ldots,m$

3) Order the b_{j1}^* from smallest to largest, and obtain $b_1^*\left(\alpha/2\right)$ and $b_1^*\left(1-\left(\alpha/2\right)\right)$
(the lower and upper percentiles of the distribution)

4) Compute $d_1 = b_1 - b_1^*\left(\alpha/2\right) \qquad d_2 = b_1^*\left(1-\left(\alpha/2\right)\right) - b_1$

5) Approximate $\left(1\text{-}\alpha\right)100\%$ CI for β_1 : $\quad b_1 - d_2 \le \beta_1 \le b_1 + d_1$