Contents

Remedial Measures (Ch. 11)

1. A brief summary of diagnostics

- Check normality of the residuals with a normal Q-Q plot and tests.
- Plot the residuals versus predicted values, versus each of the X's and, when appropriate, versus time.
- Examine the partial regression plots:
  - Use the graphics smoother to see if there appears to be a curvilinear pattern.
- Examine for outlier and influential cases:
  - Studentized deleted residuals.
  - Leverages (from the hat matrix).
  - Dffits, Cook's D, and the DFBETAS.
  - Check observations that are extreme on these measures relative to the other observations.
- Examine the VIF (i.e., 1/tolerance) for each X. If there are predictors with high VIF, do more model building:
  - Recode variables.
  - Variable selection.

2. A brief overview of remedial measures

- In Ch.3, we discussed a few remedial measures, such as transforming the data and building more complex models.
- For data with non-constant error variance: Weighted Least Squares, etc.
- For data with multicollinearity: Ridge Regression, Principle Component Analysis, Factor analysis, etc.
- For outliers and influential cases: Robust Regression, Weighted Least Squares, etc.

- For non-Normal and/or non-linear data: Generalized Linear model, Loess methods, Regression and classification tree, etc.

- For dependent/correlated errors: Time series, General Linear model (repeated measure), etc.

- Use other methods (such as bootstrap) for statistical inference.

3. Weighted Least Squares

- Recall OLS: find $b=(b_0, b_1, b_2, \dots, b_{p-1})^T$ such that
$$Q=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-(b_0+b_1X_{i1}+b_2X_{i2}+\dots+b_{p-1}X_{i,p-1})\right)^2$$
is minimized. In matrix form, $b=\left(X^TX\right)^{-1}X^TY$

- **Weight Least Squares**: Assign a weight w_i to each case, find $b_w=(b_0, b_1, b_2, \dots, b_{p-1})^T$ such that
$$Q_w=\sum_{i=1}^{n}w_i\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}w_i\left(Y_i-(b_0+b_1X_{i1}+b_2X_{i2}+\dots+b_{p-1}X_{i,p-1})\right)^2$$
is minimized.

- Let W be a diagonal matrix of weights (i.e., the weight matrix), W = diag (w_1, …, w_n). The weighted least squares estimate of β is $b_w=\left(X^TWX\right)^{-1}X^TWY$.

- To solve the issues on non-constant error variance (i.e., heteroscedasticity, heterogeneity), let $w_i=1/\sigma_i^2$.
  - Why would this work?
    - Some math/stat background

$$Y_i\sim N\left((b_0+b_1X_{i1}+b_2X_{i2}+\dots+b_{p-1}X_{i,p-1}),\quad \sigma_i^2\right)$$

$$Y_i^*=\frac{1}{\sigma_i}Y_i\sim N\left(\frac{1}{\sigma_i}(b_0+b_1X_{i1}+b_2X_{i2}+\dots+b_{p-1}X_{i,p-1}),\quad 1\right)$$

$$\begin{aligned}Q_w&=\sum_{i=1}^{n}\left(Y_i^*-\hat{Y}_i^*\right)^2\\&=\sum_{i=1}^{n}\left(\frac{1}{\sigma_i}Y_i-\frac{1}{\sigma_i}(b_0+b_1X_{i1}+b_2X_{i2}+\dots+b_{p-1}X_{i,p-1})\right)^2\\&=\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left(Y_i-(b_0+b_1X_{i1}+b_2X_{i2}+\dots+b_{p-1}X_{i,p-1})\right)^2\end{aligned}$$

- Weighted least square can be considered as an OLS on the transformed data. The transformation is built in the method, though it is not shown in the spread sheet.
- Weighted least squares estimates are the same as the MLE in the above setting.

➢ Determination of the weights.

- If σ_i^2's are known, $w_i = 1/\sigma_i^2$.

- If σ_i^2 , i = 1, 2, …, n, are unknown, find a way to estimate them. Here are a few choices:
  - Use grouped data or approximately grouped data to estimate the variance.
  - Find a relationship between the <u>squared residual</u> and another predictor variable and use this as a model for the variance. Let $\hat{v}_i$ be the estimated variance of the i-th case. Set $w_i = 1/|\hat{v}_i|$
  - Find a relationship between the <u>absolute residual</u> and another predictor variable and use this as a model for the standard deviation. Let $\hat{s}_i$ be the estimated standard deviation of the i-th case. Set $w_i = 1/(\hat{s}_i)^2$
  - Often, the weights are estimated in an iterative fashion to improve the results. This is also known as the Iterative Weighted Least Squares.

➢ See R and SPSS lab for a complete example.

➢ Weighted least squares may be used in other scenarios.

- Account for the sampling weights, or to adjust the sampling weights.
- Lower the influence from some outliers.
- Work with correlated data. (More advance. Need to estimate the correlation structure first.)

4. Ridge Regression

➢ An old idea from numerical analysis: if (X^TX) is difficult to invert (near singular) then approximate by inverting $(X^TX+\lambda I)$.

➢ Using standardized data, Y^* and X^*, $b_{Ridge} = ((X^*)^T(X^*) + \lambda I)^{-1}(X^*)^T Y^*$, where the constant λ is the "ridge parameter."

➢ Estimates of the regression coefficients are <u>biased</u> but <u>have smaller standard errors</u>.

➢ To choose the optimal λ, use a "ridge trace" plot or other estimation procedures. See example in text, p.434, 435, and in R handout.

- In SPSS, you will have to use syntax to program for ridge regression. In R, check the lm.ridge( ) function in the MASS library.

5. Robust regression

- The Basic idea is to have a procedure that is not sensitive to outliers.

- Alternatives to least squares:
  - Minimize the sum of the absolute values of the residuals: $\sum_{i=1}^{n} | Y_i - \hat{Y}_i |$.
  - Minimize the median of the squares of the residuals: $median\left(Y_i - \hat{Y}_i\right)^2$.

- Do weighted regression with weights based on residuals, and iterate.

- More computationally intensive, usually no closed-form solution and requires numerical solutions.

6. Bootstrap

- It is based on simulation and resampling.

- Sample with replacement from the data and repeatedly refit the model to get the sampling distribution of the parameter of interest.

- It is a very important theoretical development that has had a major impact on applied statistics.

7. Other remedial measures

- Principle component analysis and factor analysis.
  - Reduce multicollinearity.
  - Reduce the dimension (number of predictors) of the problem.

- Non-parametric methods:
  - Smoothing, piece-wise regression fit (loess fit).
  - Regression tree, decision tree, classification and regression tree (CART).