Miscellaneous Notes on Regression

Based on SJS and KNN

© Tingfeng Xia

Fall 2019, modified on November 8, 2019

Preface

Notes for STA302H1F fall offering, 2019 with Prof. Shivon Sue-Chee. These notes are based on the KNN and SJS text, in an aim for better understanding of the course material.

This work is licensed under a Creative Commons "Attribution-NonCommercial-ShareAlike 4.0 International" license.

Contents

1 Diagnostics and Transformations for SLR

1.1 Valid and Invalid Data

1.1.1 Residuals

One tool that we can use to validate a regression model is one or more plots of residuals (or standardized residuals). These plots will enable us to assess visually whether an appropriate model has been fit to the data no matter how many predictor variables are used.

Expected Behavior We expect that the residual graph to have no discernable pattern and centered at some value (0 in the case of standardized residual). Patterns such as curves, skewness et cetra indicates non-normal residuals. More on this in the below section.

1.1.2 Reading Residual Plots

Criterion One way of checking whether a valid simple linear regression model has been fit is to plot residuals versus x and look for patterns. If no pattern is found then this indicates that the model provides an adequate summary of the data, i.e., is a valid model. If a pattern is found then the shape of the pattern provides information on the function of x that is missing from the model.

Rationale Suppose that the true model is a straight line (which we never know) defined as

$$Y_i = \mathrm{E}\left(Y_\mathrm{i}|X_i = x_i\right) + e_i = \beta_0 + \beta_1 x_i + e_i \tag{1}$$

where

$$e_i = \text{Random error on } Y_i \quad \text{and } E(e_i) = 0$$

and we fit a regression line

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Under the assumption that our regression line is very close to the true model, i.e. $\beta_0 \approx b_0$ and $\beta_1 \approx b_1$), we see

$$\begin{aligned}
\hat{e}_i &= y_i - \hat{y}_i \\
&= \beta_0 + \beta_1 x_i + e_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\
&= \left(\beta_0 - \hat{\beta}_0\right) + \left(\beta_1 - \hat{\beta}_1\right) x_i + e_i \\
&\approx e_i
\end{aligned}$$

which means that our residuals resembles the random error!

1.2 Regression Diagnostics

Catagorization

- **X-Direction Outlier, i.e. Leverage Point**: Away from the bulk of data in the x-direction.

- **Good:** Not much chaneg after removing the data point, i.e. the data point originally was quite close to the regression line although away from the bulk of data in the x direction. *"A good leverage point is a leverage point which is NOT also an outlier."*
- **Bad, Influential Point:** If its Y-value does not follow the pattern set by the other data points, i.e. a bad leverage point is a leverage point which is also an outlier.

- **Y-Direction Outlier** Trait: large residuals

1.2.1 Leverage Point

Defining The Hat The hat came from yet another representation of the $\hat{y}_i$. Recall that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $\hat{\beta}_1 = \sum_{j=1}^n c_j y_j$ and $c_j = \frac{x_j - \bar{x}}{SXX}$. Then we have

$$\begin{aligned}\hat{y}_i &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\ &= \frac{1}{n}\sum_{j=1}^n y_j + \sum_{j=1}^n \frac{(x_j - \bar{x})}{SXX} y_j (x_i - \bar{x}) \\ &= \sum_{j=1}^n \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX}\right] y_j = \sum_{j=1}^n h_{ij} y_j\end{aligned}$$

where we define

$$h_{ij} = \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX}\right]$$

Property of The Hat Recall that $\sum_{j=1}^n [x_j - \bar{x}] = n\bar{x} - n\bar{x} = 0$, then

$$\sum_{j=1}^n h_{ij} = \sum_{j=1}^n \left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX}\right] = \frac{n}{n} + \frac{(x_i - \bar{x})}{SXX}\sum_{j=1}^n [x_j - \bar{x}] = 1$$

Thus,

$$\hat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j \quad \text{where } h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

Defining Leverage The term $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$ above is commonly known as the leverage of the ith data point. Notice the following in the definition of the leverage h_{ii}

- The second term measures the proportion, in terms of squared deviation in x-direction over sum of square of total devaition in x-direction, of the i-th data point's deviation. When the second term tends to 1, meaning that i-th data point is some extreme outlier in the x-direction, then h_{ii} would close to one, signifiying the 'leverage'-ness.
- Recall that $\sum_{j=1}^n h_{ij} = 1$, then when $h_{ii} \cong 1$, $h_{ij} \to 0$ and

  $$\hat{y}_i = 1 \times y_i + \text{ other terms } \cong y_i$$

  which means $\hat{y}_i$ will be very close to y_i, regardless of the rest dataset.

- A point of high leverage (or a leverage point) can be found by looking at just the values of the x's and not at the values of the y's

Average of Leverage For simple linear regression,

$$\text{average}\,(h_{ii}) = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{2}{n}$$

Identifying Leverage Rule: x_i is a high leverage (i.e., a leverage point) in a SLR model if

$$h_{ii} > 2 \times \text{average}\,(h_{ii}) = 2 \times 2/n = 4/n$$

Dealing with 'Bad' Leverage

- **Remove invalid data points;** Question the validity of the data points corresponding to bad leverage points, that is: Are these data points unusual or different in some way from the rest of the data? If so, consider removing these points and refitting the model without them.
- **Fit a different regression model;** Question the validity of the regression model that has been fitted, that is: Has an incorrect model been fitted to the data? If so, consider trying a different model by including extra predictor variables (e.g., polynomial terms) or by transforming Y and/or x (which is considered later in this chapter).

1.2.2 Standardized Residuals

Problem of Non-constant Variance Recall that

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX}$$

and (we will show this later)

$$\text{Var}\,(\hat{e}_i) = \sigma^2\,[1 - h_{ii}]$$

which is indeed non-constant for different data points. When $h_{ii} \cong 1$ (h_{ii} is very close to 1), the i-th data point is a leverage point and

$$\text{Var}(\hat{e}_i) = \sigma^2\,[1 - h_{ii}] \approx 0 \quad \text{and} \quad \hat{y}_i \cong y_i$$

The above results intuitively makes sense: When i-th data point is a leverage, $\hat{e}_i$ will be small and it does not vary much (data point close to the estimated regression line).

Derivation of Residual Variance (Not Important) Recall that

$$\hat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j \quad \text{where } h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX}$$

Then,

$$\hat{e}_i = y_i - \hat{y}_i = y_i - h_{ii}y_i - \sum_{j\neq i} h_{ij}y_j = (1 - h_{ii})\,y_i - \sum_{j\neq i} h_{ij}y_j$$

Hence,

$$\begin{aligned}
\operatorname{Var}(\hat{e}_i) &= \operatorname{Var}\left((1-h_{ii})\,y_i - \sum_{j\neq i} h_{ij}y_j\right) \\
&= (1-h_{ii})^2\,\sigma^2 + \sum_{j\neq i} h_{ij}^2\sigma^2 \\
&= \sigma^2\left[1 - 2h_{ii} + h_{ii}^2 + \sum_{j\neq i} h_{ij}^2\right] \\
&= \sigma^2\left[1 - 2h_{ii} + \sum_{j} h_{ij}^2\right]
\end{aligned}$$

Notice that

$$\begin{aligned}
\sum_{j=1}^{n} h_{ij}^2 &= \sum_{j=1}^{n}\left[\frac{1}{n} + \frac{(x_i-\bar{x})\,(x_j-\bar{x})}{SXX}\right]^2 \\
&= \frac{1}{n} + 2\sum_{j=1}^{n}\frac{1}{n}\times\frac{(x_i-\bar{x})\,(x_j-\bar{x})}{SXX} + \sum_{j=1}^{n}\frac{(x_i-\bar{x})^2\,(x_j-\bar{x})^2}{SXX^2} \\
&= \frac{1}{n} + 0 + \frac{(x_i-\bar{x})^2}{SXX} \\
&= h_{ii}
\end{aligned}$$

So,

$$\operatorname{Var}(\hat{e}_i) = \sigma^2\left[1 - 2h_{ii} + h_{ii}\right] = \sigma^2\left[1 - h_{ii}\right]$$

and

$$\operatorname{Var}(\hat{y}_i) = \operatorname{Var}\left(\sum_{j=1}^{n} h_{ij}y_j\right) = \sum_{j\neq i} h_{ij}^2\operatorname{Var}(y_j) = \sigma^2\sum_{j} h_{ij}^2 = \sigma^2 h_{ii}$$

Overcome with Standardization The above problem of each $\hat{e}_i$ having different variances could be overcame by standardizing the residuals. The i-th standardized residual is defined as (notice that the $s = \hat{\sigma}$ is the estimated variance in the SLR settings)

$$r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}} \quad \text{where } s = \sqrt{\frac{1}{n-2}\sum_{j=1}^{n}\hat{e}_j^2}$$

Advantages of Standardization

- When points of high leverage exist, instead of looking at residual plots, it is generally more informative to look at plots of standardized residuals since plots of the residuals will have nonconstant variance even if the errors have constant variance.

- When points of high leverage do not exist, there is gener- ally little difference in the patterns seen in plots of residuals when compared with those in plots of standardized residuals.
- The other advantage of standardized residuals is that they immediately tell us how many estimated stand- ard deviations any point is away from the fitted regression model.

Recognizing Outliers Using Standardized Residuals

- An **outlier** is a point whose standardized residual falls outside the interval from -2 to 2, i.e. $|r_i| > 2$
- A **Bad Leverage Point** is a leverage point whose standardized residual falls outside the interval from -2 to 2, i.e. $|r_i| > 2 \wedge h_{ii} > \frac{4}{n}$
- A **Good Leverage Point** is a leverage point whose standardized residual falls inside the interval from -2 to 2, i.e. $|r_i| \leq 2 \wedge h_{ii} > \frac{4}{n}$
- **Dealing with large datasets:** In this case, we should change the above criterion to $r_i| > 4$ and $r_i| \leq 4$ respectively. This is to give allowance for more occurence of rare events in a large data set.

Correlation Between Residuals Even if the errors are independent (homogeneous), i.e. $e_i \perp\!\!\!\perp e_j$ $(i \neq j)$, the residuals are still correlated. It can be shown that the covariance and the correlation is given by

$$\text{Cov}\left(\hat{e}_i, \hat{e}_j\right) = -h_{ij}\sigma^2 (i \neq j)$$

$$\text{Corr}\left(\hat{e}_i, \hat{e}_j\right) = \frac{-h_{ij}}{\sqrt{\left(1 - h_{ii}\right)\left(1 - h_{ij}\right)}} (i \neq j)$$

Such correlation could be safely ignored in practice. They are usualy given raise by inherent correlation such as data collected over time.

1.2.3 Recommendations for Handling Outliers & Leverage

We have discussed multiple ways of assessing outliers and talked about the way to deal with them by removing them. However, it is not always a good idea to delete them for the following reasons:

- Points should not be routinely deleted from an analysis just because they do not fit the model. Outliers and bad leverage points are signals, flagging potential problems with the model.
- Outliers often point out an important feature of the problem not considered before. They may point to an alternative model in which the points are not an outlier. In this case it is then worth considering fitting an alternative model.

1.2.4 Influence of Certain Cases

It can sometimes be the case that certain data points in a data set are drastically controlling the entire regression model (the model has payed too much attention to them). We now decelop methods where we measure the "importance" of a data point.

Cook's Distance First, define (recall if already defined) the following notation

- $\hat{y}_{j(i)}$ means the the fitted value of the j-th data point on the regression line obtained by removing the i-th case.
- $S^2 = \frac{1}{n-2}\sum_{j=1}^{n}\hat{e}_j^2$ is the variance (Original MSE) of the total regression model.
- $r_i = \frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}$ where $s = \sqrt{\frac{1}{n-2}\sum_{j=1}^{n}\hat{e}_j^2}$

Then, the Cook's Distance of the i-th datapoint is given by

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_{j(i)} - \hat{y}_j\right)^2}{2S^2} = \frac{r_i^2}{2}\frac{h_{ii}}{1-h_{ii}}$$

where we should note that D_i may be large due to large r_i, or large h_{ii} or both.

Rule: Cook's Distance

- A point is noteworthy if
$$D_i > \frac{4}{n-2}$$
- In practice, look for gaps in the values of Cook's Distance and not just whether one value exceeds tyhe suggested cut off.

1.2.5 Normality of the Errors

The assumption of normal errors is (especially) needed in small samples for the validity of t-dist based tests and inferences. This assumption is generally checked by looking at the distribution of the residuals or standardized residuals. Recall that the i-th least squares residuals is given by $\hat{e}_i = y_i - \hat{y}_i$. We will now show $\hat{e}_i = e_i - \sum_{j=1}^{n} h_{ij}e_j$. First, in the derivation we will need these two facts

$$\sum_{i=1}^{n} h_{ij} = 1$$

and

$$\sum_{j=1}^{n} x_j h_{ij} = \sum_{j=1}^{n}\left[\frac{x_j}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})x_j}{SXX}\right] = \bar{x} + \frac{(x_i-\bar{x})SXX}{SXX} = x_i$$

We then proceed as follows

$$\begin{aligned}
\hat{e}_i = y_i - \hat{y}_i &= y_i - h_{ii}y_i - \sum_{j \neq i} h_{ij}y_j \\
&= y_i - \sum_{j=1}^{n} h_{ij}y_j \\
&= \beta_0 + \beta_1 x_i + e_i - \sum_{j=1}^{n} h_{ij}\left(\beta_0 + \beta_1 x_j + e_j\right) \\
&= \beta_0 + \beta_1 x_i + e_i - \beta_0 - \beta_1 x_i - \sum_{j=1}^{n} h_{ij}e_j \\
&= e_i - \sum_{j=1}^{n} h_{ij}e_j
\end{aligned}$$

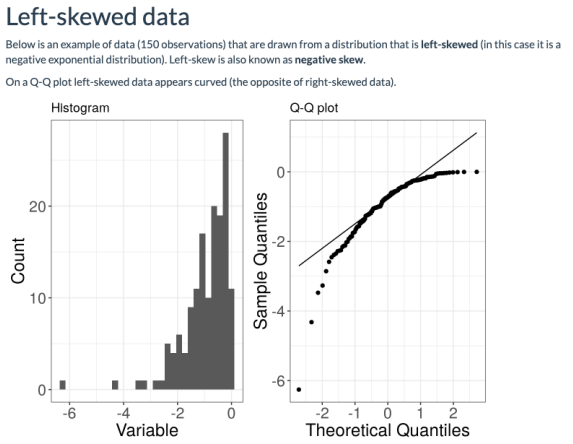
■

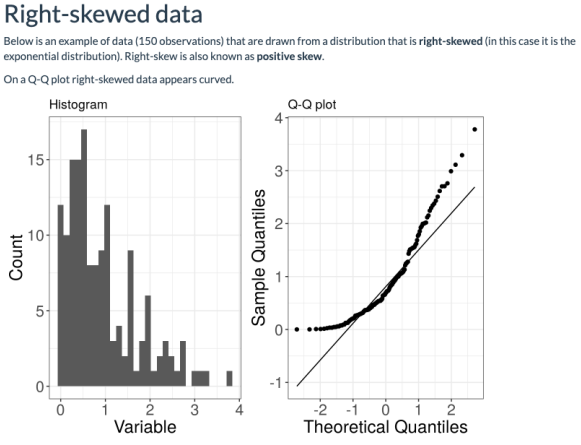
The above result showed that the i-th least squares residual is equal to e_i minus a weighted sum of all the e's. There are two cases to consider,

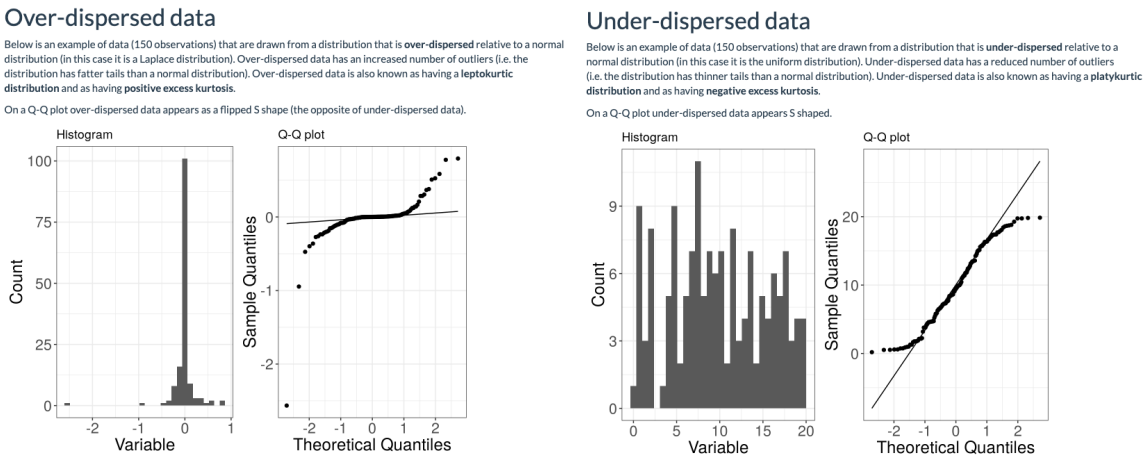
- In small to moderate samples, the second term could dominate the first and first and the residuals can look like they come from a normal distribution even if the errors do not.
- When n is large, the second term in the derived result (thistle colored) has a much smaller variance than that of the first term and as such the first term dominates the last equation.

Conclusion: For large samples, the residuals can be used to assess normality of the errors.

Assessment Using Normal Q-Q A normal probability plot of the standardized residuals is obtained by plotting the ordered standardized residuals on the vertical axis against the expected order statistics from a standard normal distribution on the horizontal axes. If the resulting plot produces points "close" to a straight line then the data are said to be consistent with that from a normal distribution. On the other hand, departures from linearity provide evidence of non-normality.







1.2.6 Constant Variance

2 Weighted Least Square Regression

2.1 Motivation and Set-Up

Consider the straight line (simple) linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{where} \quad e_i \sim N(0, \frac{\sigma^2}{w_i})$$

For the weight w_i, we should note the following

- $w_i \to \infty \implies Var(e_i) \to 0$. In this case, the estimates of the regression parameters β_0, β_1 should be such that the fitted line at x_i should be very close to y_i. (Small variance means more strict in terms of deviation from the regressin line, corresponding to a larger emphasis on the i-th data point.)
- If w_i is some small value, then the variance of the i-th data point would be quite large. In this case, we have a loose restriction of the deviation of the i-th data point from the regression line meaning that litte emphasis is taken for this data point.
- $w_i \to 0 \implies Var(e_i) \to \infty$. In this case, we have the variance tending to infinity. Meaning that there is absolutely no restriction/emphasis on the i-th data point and it could be simply removed from the set.

We define the cost function, WRSS as

$$\text{WRSS} = \sum_{i=1}^{n} w_i \left(y_i - \hat{y}_{W_i}\right)^2 = \sum_{i=1}^{n} w_i \left(y_i - b_0 - b_1 x_i\right)^2$$

and the estimators $\mathbf{b} = [b_0, b_1]^T$ are derived using MLE.

Intuition behind WRSS This cost function may seem wierd at first glance, but it intuitively makes sense. Notice that when w_i is large, the i-th lost term $w_i \left(y_i - \hat{y}_{W_i}\right)^2$ is payed more emphasis on. On the contrary, when $w_0 \to 0$, the term $\to 0$. (Indeed, when Variance of the term $\to \infty$ we just neglect it.)

2.2 Deriving ML Estimators

Derivatives

$$\frac{\partial \text{WRSS}}{\partial b_0} = -2\sum_{i=1}^{n} w_i \left(y_i - b_0 - b_1 x_i\right) = 0 \tag{2}$$

$$\frac{\partial \text{WRSS}}{\partial b_1} = -2\sum_{i=1}^{n} w_i x_i \left(y_i - b_0 - b_1 x_i\right) = 0 \tag{3}$$

Normal Equations Obtained from rearranging the above equations, we will call them Normal Eq1 and Normal Eq2 respectively for later reference.

$$\sum_{i=1}^{n} w_i y_i = b_0 \sum_{i=1}^{n} w_i + b_1 \sum_{i=1}^{n} w_i x_i \tag{4}$$

$$\sum_{i=1}^{n} w_i x_i y_i = b_0 \sum_{i=1}^{n} w_i x_i + b_1 \sum_{i=1}^{n} w_i x_i^2 \tag{5}$$

Rearranging Use Normal Eq1 $\times \sum_{i=1}^{n} w_i x_i$ and Normal Eq2 $\times \sum_{i=1}^{n} w_i$

$$\sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i = b_0 \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i + b_1 \left(\sum_{i=1}^{n} w_i x_i\right)^2 \tag{6}$$

$$\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i y_i = b_0 \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i + b_1 \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2 \tag{7}$$

WLS Slope Estimator [1]

$$\hat{\beta}_{1W} = \frac{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i y_i - \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2 - \left(\sum_{i=1}^{n} w_i x_i\right)^2} \tag{8}$$

$$= \frac{\sum_{i=11}^{n} x_i \left(x_i - \bar{x}_W\right)\left(y_i - \bar{y}_W\right)}{\sum_{i=1}^{n} w_i \left(x_i - \bar{x}_W\right)^2} \tag{9}$$

WLS Intercept Estimator

$$\hat{\beta}_{0W} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} - \hat{\beta}_{1W} \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \bar{y}_w - \hat{\beta}_{1W} \bar{x}_W \tag{10}$$

3 Multiple Linear Regression

[1] Note that $\bar{x}_W = \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i$ and $\bar{y}_W = \sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i$