Lecture 15 Diagnostics and Remediation

Objectives

- Identifying and dealing with Unusual and Influential observations.
- Identifying and dealing with non linearity, non constant error and non normality.

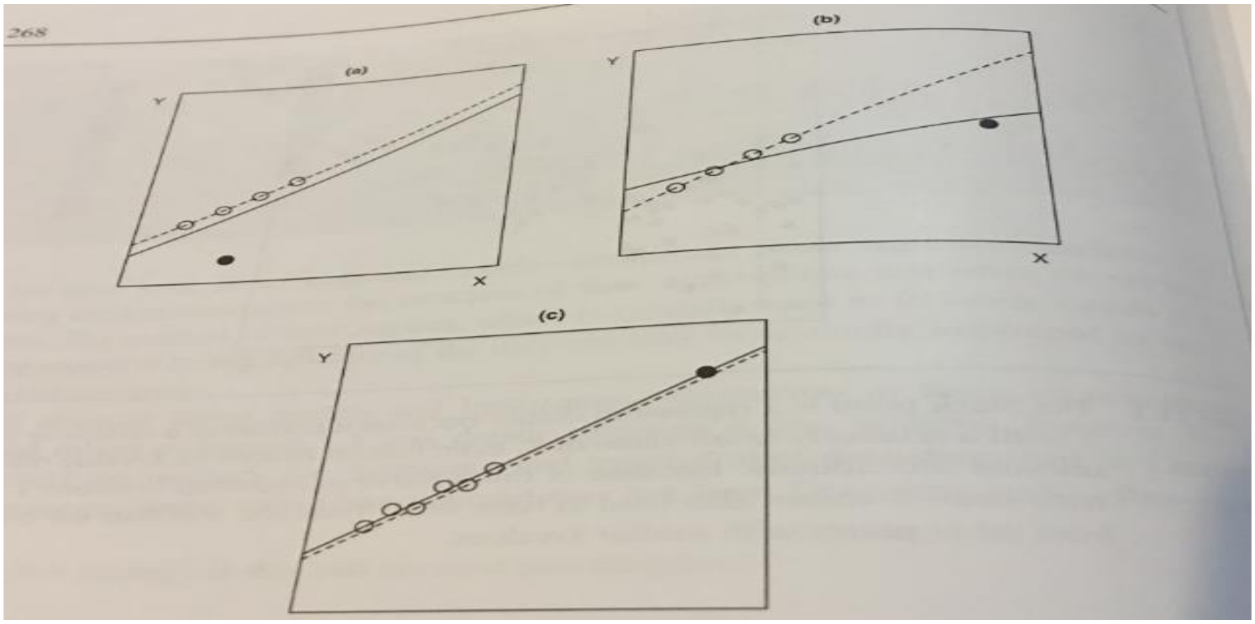
Introduction

- The method of least squares used to fit the model to the data gets affected by the structure of real life data that does not follow the assumptions of Linear Models (Linearity, Constant Variance, Normality and Uncorrelated Errors).
- Another major problem that is encountered is the problem of unusual and influential data.
- The diagnostic and corrective actions require identifying and remediating these.
- If the diagnosis and remediation is done subsequent to fitting the model to the data it is called a post fit.
- If we actually looked at the structure of the data prior to the fit and if we fixed these issues by transforming etc then the post fit will encounter less problems.

Outliers, Leverage and Influence

The value of the outlier is the value of Y response variable conditional to the X response variable.

$Y \mid X = \alpha + \beta X + \varepsilon$ ……………………………………………………1



- For b) as compared to a) the X value is much higher therefore it exerts a higher Leverage on the slope and intercept coefficient. Deleting the outlier for b) will affect the coefficients significantly.
- The discrepancy also affects the coefficient. The discrepancy: distance of the outlier from the line of least squares. The discrepancy for b) is also significant.
- Heuristically the Influence of an outlier on the coefficients is proportional to the Leverage and discrepancy.

$$Influence\_on\_Coefficients = Leverage * Discrepancy \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$$

Leverage:

Is captured by the weight h_{ij} of the observation Y_i on the fitted value $\hat{Y}_j$

Without the rigorous proof the weight h_{ii} also named h_i is the leverage that Y_i exerts on all fitted values. These are called the hat values. The hat values are bounded by 1/n and 1.

$$h_i = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \quad \text{where} \quad \overline{h} = \frac{k+1}{n} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3$$

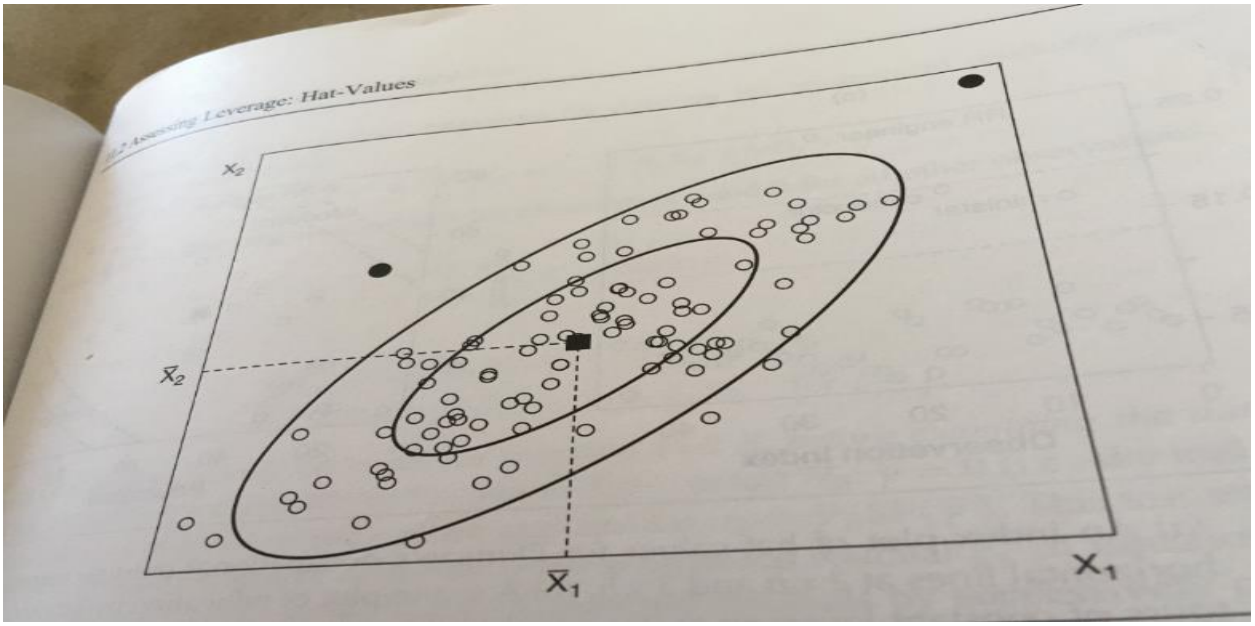
k is the number of regressors excluding the constant.

n is the number of observations.

Simple Regression h_i depicts the distance from the mean of X

Multiple Regression measures the distance from the Centroid (point of means of X).

For the Davis Data set we plotted the reported weight and measured weight. The hat value of 12th subject was .714 wheras the $\bar{h}$ mean was .0219. Therefore the 12th observation has a high leverage.



Detecting Outliers, Testing for Outliers:

Higher discrepancy values have large residuals E_i .

The standardized residual can be calculated by the formula:

$$E_i' = \frac{E_i}{S_E\sqrt{1-h_i}} \quad \text{where } S_E = \sqrt{\frac{\sum E_i^2}{n-k-1}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 4$$

h_i is the leverage that can be calculated by the equation 3

This measure is not used since numerator and denominator are not independent. E_i' is therefore does not follow t distribution

To remediate this the model can be refitted by deleting the ith observation , calculating $S_{E(-1)}$ as an estimate of σ_{ε} using n-1 observations and then calculating studentized residual :

$$E_i^* = \frac{E_i}{S_{E(-1)}\sqrt{1-h_i}}$$

Measuring Influence

Influence on regression coefficient is captured by discrepancy and leverage.

Influence can be measured by the impact on each coefficient of deleting each observation in turn.

$D_{ij} = B_j - B_{j(-1)}$ for $i = 1,2,3....n \quad j = 0,1,2.....k$

B_j Least square coefficient for all the data

$B_{j(-1)}$ Least square coefficient calculated with the ith observation deleted from the data.

This can be value can be scaled by dividing by the S_E of the deleted Standard Error.

$$D_{ij}^* = \frac{D_{ij}}{SE_{(-i)}(B_j)}$$

The problem of this measure is the large number of D_{ij} or D_{ij}^* which is n(k+1)

D_{ij} is often termed: $DFBETA_{ij}$

D_{ij}^* is often termed: $DFBETAS_{ij}$

An alternative was provided by Cook. The distance formula provided by Cook was

$$D_i = \frac{E_i^{'2}}{k+1} * \frac{h_i}{1-h_i}$$

The first term captures the discrepancy wheras the second term captures the leverage.

Belsley provides another measure for influence:

$$DFFITS = E_i^* \sqrt{\frac{h_i}{1-h_i}}$$

For the Davis Data the outlier observation was the 12th observation and it was a female data value

Cook's D_{12}=85.9(the nearest value is D_{115}=.085)

$DFFITS_{12}$=-38.4(next nearest $DFFITS_{115}$=.603)

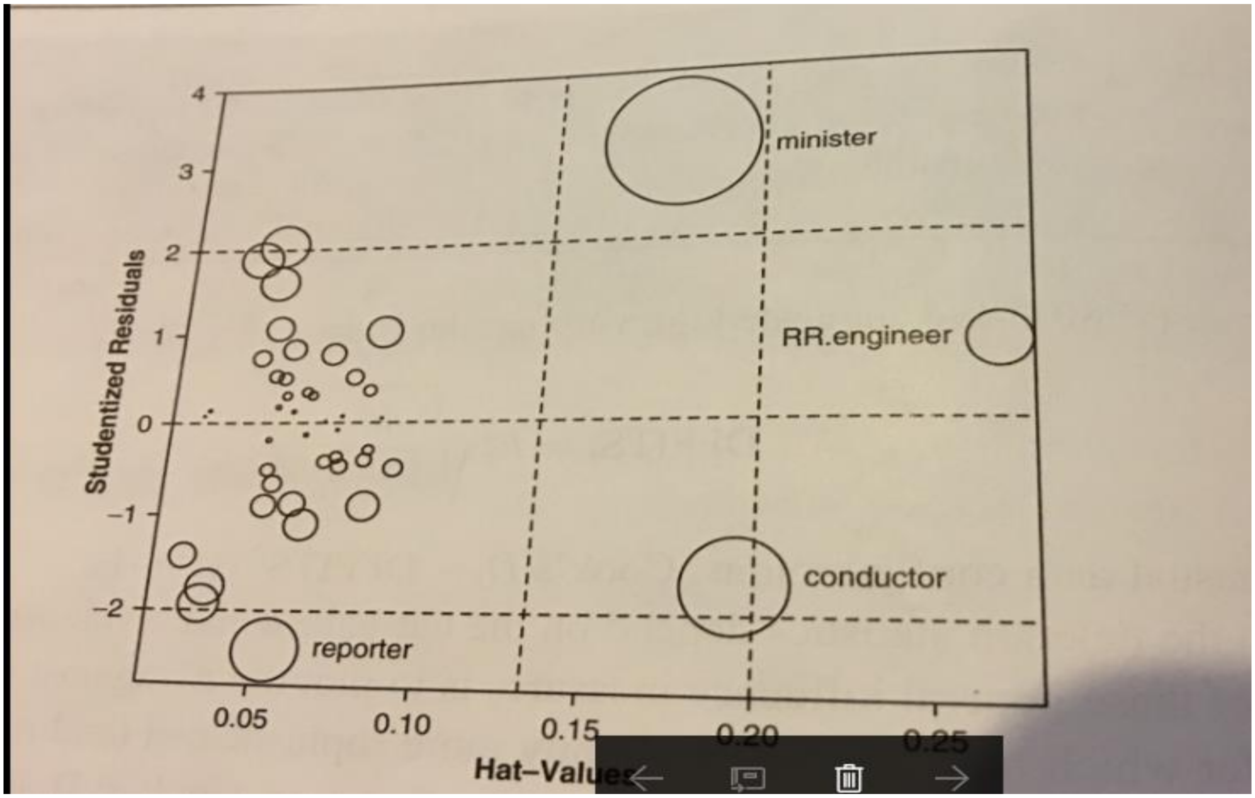
$DFBETAS_{0,12}$=$DFBETAS_{1,12}$=0 These are the coefficient for males and the outlier was in the female data. So the male intercept and slope coefficient are not affected.

For women $DFBETAS_{2,12}$=20.0 $DFBETAS_{3,12}$=-24.8

The dummy coefficient B_2 gets affected and so does the interaction coefficient B_3.

Bubble plots

Plots Hat values with the studentized residuals with the area of the bubbles proportional to the cook's distance.



The outlier values should not be deleted without making a determination of why the outlier was actually present in the data. Was it a data entry mistake or was the data unique due to some reason. The outliers will give more insight into the data and sometimes we might add some explanatory variables to improve the model for handling outliers.

Diagnosing Non Normality Non Constant Error Variance and Non Linearity

These are the initial assumptions that we imposed to implement Regression.

If these are violated then we do not obtain reliable estimates.

Non Normally Distributed Errors

The assumptions of normally distributed error is an important one.

By central limit theorem for a large enough sample the Least square estimation of coefficients and CI is robust even though the errors may not be normally distributed.

The least square is robust but the efficiency of the unbiased estimators is only achieved if the errors are normally distributed.

Efficiency is determined by Minimum Variance Unbiased Estimator: MVUE

The efficiency of least square estimators decreases substantially specially for heavy tails distribution because this gives rise to outliers.

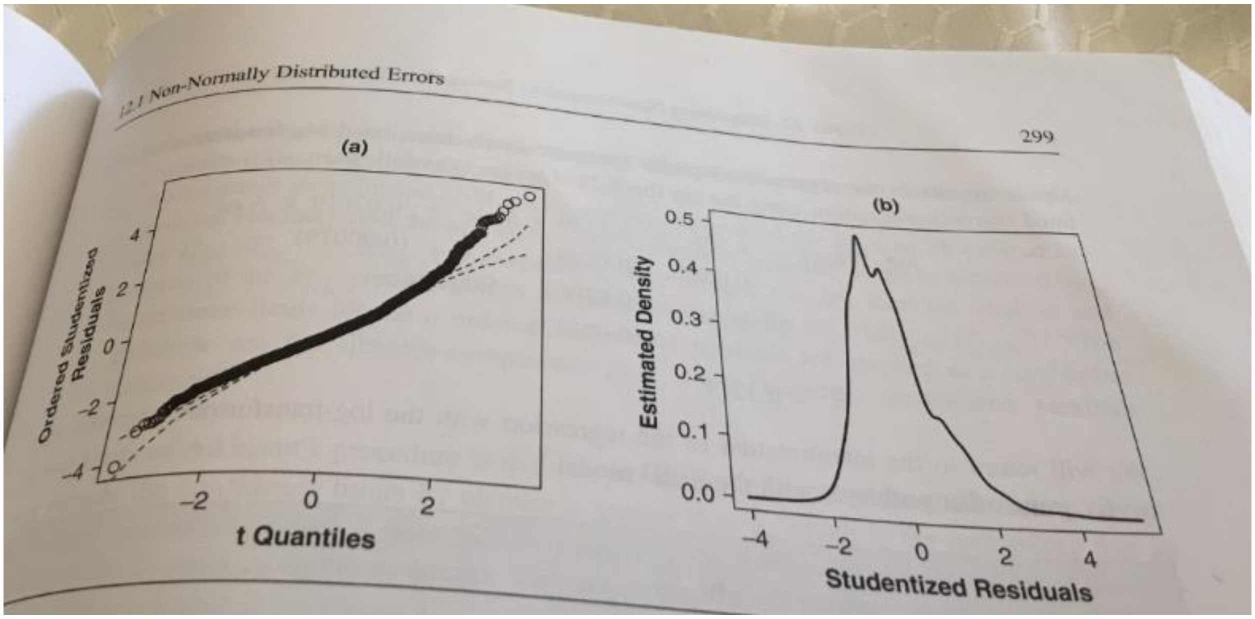
Gauss Markov's theorem states that Least Square estimators are most efficient unbiased estimators depends on the assumption of linearity, constant error variance and independence but not normality. These restrictions are able to formulate the coefficient standard error but it is not is not very strong given that the least square regression is affected by the heavy tail error distribution.

Skewed data usually generate errors in the direction of the skew. Mean therefore is not a good measure of central tendency. To deal with skewness transformations have to be used.

Multimodal error distribution is due to omission of some explanatory variable that divides the data into natural groups.

Various tests can be performed to test non normal errors but graphical displays are easier to use to pinpoint the problem.

QQ can be used. We can use the Studentized residuals with the t distribution Residuals.



QQplot show us the skewness and the tail behavior. Here the data is clearly right skewed. The dotted line shows the 95% confidence interval with assumption that the errors are normally distributed.

The density plot of the residuals gives us the indication that there might be 2 modes.

As we have read previously right skew of a distribution can be remediated by transforming down the ladder of powers. Log transformation is a good one and so does cube root. The regression results for log transformation or cube roots are similar so either or can be used.

Non Constant Error Variance

One of our assumptions was that the error variance (also the same as the variance of response variable) has to be constant.

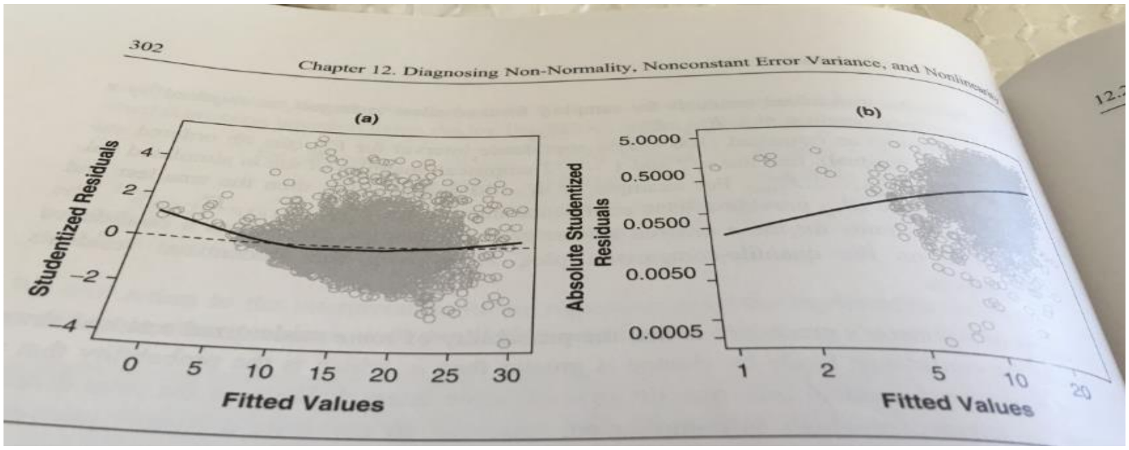
$$V(\varepsilon) = V(Y \mid x_1, x_2 \ldots\ldots\ldots x_k) = \sigma_\varepsilon^2$$

Constant error variance is often called homoscedasticity. If this assumption is violated by the data the efficiency of the least squares unbiased estimator is compromised. The formulas that we apply for the standard errors for the regression coefficients are inaccurate in this context. The degree of inaccuracy is determined by the extent to which the error variances differ, the sample size as well as distribution of X values in the data.

Residual Plots:

The general trend that we might observe is that the error variance increases as the expectation of Y grows. This can be detected by plotting residuals (usually studentized: shows spread pattern more clearly) with the fitted values. There also might be a systematic relationship between the error variance and a particular X. This trend can be detected by plotting residuals against each X.

For plots given below we can clearly see that the spread is increasing with the level of response.



To remediate this the transformation down the ladder of plots has to be applied. The slope of the spread level plot for the fitted value is b=.9994 therefore the power transformation is 1-.9994=.0006.

We can apply log transformation. This will make the error variance more constant.

This is the same transformation that we applied when we had non normally distributed errors.

Transformations can both normalize the error distribution and make the error variance more constant and therefore might sometimes linearlize the relationship between X and Y.

Linearity should be checked despite it being corrected in some cases by the transformation.

A rough rule is that non constant error variance seriously degrades the least square estimator only when the ratio of the largest to smallest variance is 10 or more (or conservatively 4 or more).

Non Linearity

The assumption of linearity has to be satisfied for the application of regression. This entails that $E(\varepsilon)=0$ expectation of error is zero.

If two explanatory variables are supposed to have an additive effect instead interact then the average error is not zero for all combinations of X values.

If the non linearity is slight the model can be an approximation for the regression surface $E(Y \mid X_1, X_2 .....X_k)=0$ otherwise the model is fallacious.

Just plotting each X explanatory variable against response variable Y does not provide a holistic picture because we are generally interested in the partial relationship of Y and each individual X (keeping all other Xs constant) instead of the relationship of Y and each individual X(ignoring all the other Xs).

Plotting residuals against each and smoothened by a non parametric regression smoother helps in the in checking non linearity.

The problem in this method is that method of least squares ensures that the residuals are linearly uncorrelated with each X.

This causes the residual to not distinguish between monotone and non monotone nonlinearity.

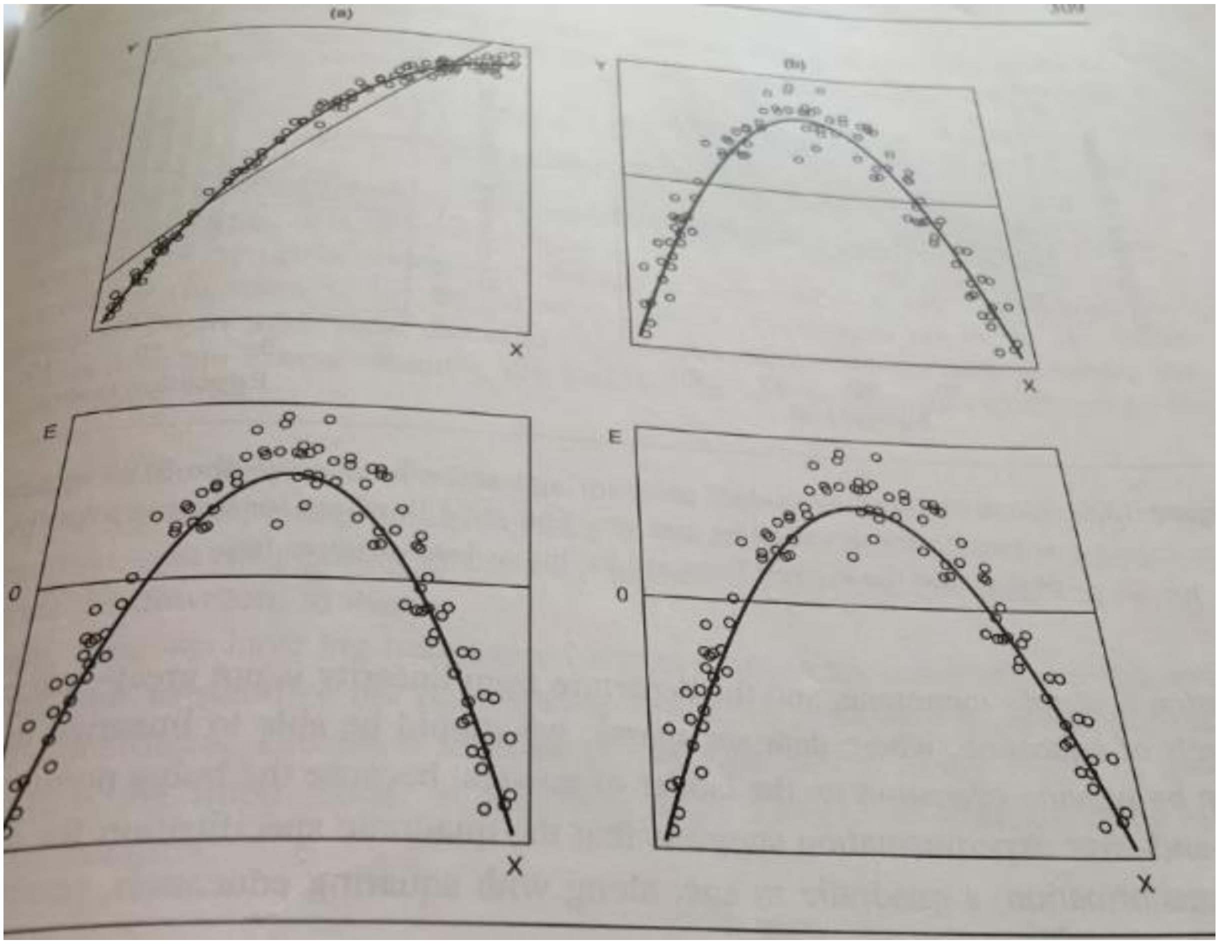
Monotone nonlinearity is corrected by simple transformations

$$Y = \alpha + \beta\sqrt{X} + \varepsilon$$

Wheras the non monotone might be linearized by a quadratic regression like

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Residuals plot therefore provide more intuition.


(a)
Y
X
(b)
Y
X
E
0
X
E
0
X