Chapter 3 – Diagnostics and Remedial Measures

Diagnostics

- Procedures to determine appropriateness of the model and check assumptions used in the standard inference
- If there are violations, inference and model may not be reasonable thereby resulting in faulty conclusions
- Always check before any inference!!!!!!!!
- Procedures involve both **graphical methods** and **formal statistical tests**

3.1 Diagnostics for the Predictor Variable (*X*)

Levels of the independent variable, particularly in settings where the experimenter does not control the levels, should be studied. Problems can arise when:

- One or more observations have *X* levels far away from the others. Influential points.
- When data are collected over time or space, *X* levels that are close together in time or space are "more similar" than the overall set of *X* levels

Useful plots of *X* levels include: dot plot, histograms, box-plots, stem-and-leaf diagrams, and sequence plots (versus time order). **See Figure 3.1 for these plots on page 101.** Also, a useful measure is simply the *z*-score for each observation's *X* value. We will later discuss remedies for these problems in Chapter 10.

Look at the first part of R codes: Chapter3_Graphs_BrownTest.R

3.2 Residuals

"True" Error Term: $\varepsilon_i = Y_i - E\{Y_i\} = Y_i - (\beta_0 + \beta_1 X_i)$

Observed Residual: $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$

- The assumption on the "true" error terms: they are independent and normally distributed with mean 0, and variance σ^2 ($\varepsilon_i \sim N(0,\sigma^2)$ and they are independent).
- The residuals have mean 0, since they sum to 0, but they are not independent since they are based on the fitted values from the same observations, but as *n* increases, this becomes less important.
- Ignoring the nonindependence for now, we have, concerning the residuals $(e_1, \ldots, e_n)$:

$$\bar{e} = \frac{\sum e_i}{n} = \frac{0}{n} = 0 \qquad s^2\{e_i\} = \frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{\sum (e_i - 0)^2}{n-2} = \frac{\sum e_i^2}{n-2} = MSE$$

Semistudentized Residuals

We are accustomed to standardizing random variables by centering them (subtracting off the mean) and scaling them (dividing through by the standard deviation), thus creating a *z*-score.
While the theoretical standard deviation of e_i is a complicated function of the entire set of sample data (we will see this after introducing the matrix approach to regression), we can approximate the standardized residual as follows, which we call the **semistudentized residuals**:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

In large samples, these can be treated approximately as *t*-statistics, with *n*-2 degrees of freedom.

Departures from the Simple Linear Regression Model with Normal Errors:

1. The regression function is not linear
2. The error terms do not have constant variance
3. The error terms are not independent
4. The model fits all but one or a few outlier observations
5. The error terms are not normally distributed
6. One or several important predictor variables have been omitted from the model

3.3 Diagnostic Plots for Residuals (or Seminstudentized residuals)

1. Plot of residuals against predictor variable
2. Plot of absolute or square residuals against predictor variable
3. Plot residuals against fitted values
4. Plot of residuals against time or other sequence
5. Plots of residuals against omitted predictor variables
6. Box plot of residuals
7. Normal probability plot of residuals QQ plot

Linear Relationship between E{*Y*} and *X*

Plot the residuals versus either *X* or the fitted values. This will appear as a random cloud of points centered at 0 under linearity, and will appear U-shaped (or inverted U-shaped) if the relationship is not linear.

Errors have Constant Variance

Plot the residuals versus *X* or the fitted values. This should appear as a random cloud of points, centered at 0, if the variance is constant. If the error variance is not constant, this may appear as a funnel shape.

Errors are Independent (When Data Collected Over Time)

Plot the residuals versus the time order (when data are collected over time). If the errors are independent, they should appear as a random cloud of points centered at 0. If the errors are positively correlated they will tend to approximate a smooth (not necessarily monotone) functional form.

Model Fits for All Observations

Plot Residuals versus fitted values. As long as no residuals stand out (either much higher or lower) from the others, the model fits all observations. Any residuals that are very extreme, are evidence of data points that are called *outliers*. Any outliers should be checked as possible data entry errors. We will cover this problem in detail in Chapter 9.

Normally Distributed Errors

Distribution plots, such as box plot, histogram, dot plot, or stem-and-leaf plot of the residuals are helpful for detecting gross departure from normality.

Alternatively, a **normal probability plot** can be obtained as follows. Here each residual is plotted against its expected value under normality. A plot that is nearly linear suggests agreement with normality (**Figure 3.9 on Page 112**).

1. Order the residuals from smallest (large negative values) to largest (large positive values). Assign the ranks as *k*.
2. Compute the percentile for each residual: $\frac{k-0.375}{n+0.25}$
3. Obtain the *z* value from the standard normal distribution corresponding to these percentiles: $z\left(\frac{k-0.375}{n+0.25}\right)$
4. Multiply the *z* values by $s=\sqrt{MSE}$ these are the “expected” residuals for the k^{th} smallest residuals under the normality assumption
5. Plot the observed residuals on the vertical axis versus the expected residuals on the horizontal axis. This should be approximately a straight line with slope 1.

Look at the second part of R codes: Chapter3_Graphs_BrownTest.R

Check Figure 3.9

No Predictors Have Been Omitted

Plot residuals versus omitted factors, or against *X* seperately for each level of a categorical omitted factor. If the current model is correct, these should be random clouds of points centered at 0. If patterns arise, the omitted variables may need to be included in model (Multiple Regression).

3.4 – 3.6 Tests Involving Residuals

Several of the assumptions stated above can be formally tested based on statistical tests.

Normally Distributed Errors (Correlation Test)

Using the expected residuals (denoted e_i*) obtained to construct a normal probability plot, we can obtain the correlation coefficient between the observed residuals and their expected residuals under normality: $r_{ee*} = \frac{\sum ee*}{\sqrt{\sum e^2 \sum (e*)^2}}$

The test is conducted as follows:

- H_0 :Error terms are normally distributed
- H_A :Error terms are not normally distributed
- *Test Statistic (TS)*: r_{ee*}
- *Rejection Region (RR)*: $r_{ee*} \leq$ Tabled values in Table B.6, Page 673 (indexed by α and *n*)

Note this is a test where we do not wish to reject the null hypothesis. Another test that is more complex to manually compute, but is automatically reported by several software packages is the Shapiro-Wilks test. It's null and alternative hypotheses are the same as for the correlation test, and *P*-values are computed for the test.

Errors have Constant Variance (Modified Levene Test) Brown-Forsythe Test:

There are several ways to test for equal variances. One simple (to describe) approach is a modified version of Levene's test, which tests for equality of variances, without depending on the errors being normally distributed. Recall that due to Central Limit Theorems, lack of normality causes us no problems in large samples, as long as the other assumptions hold. The procedure can be described as follows:

1. Split the data into 2 groups, one group with low *X* values containing n_1 of the observations, the other group with high *X* values containing n_2 observations ($n_1+n_2=n$).

2. Obtain the medians of the residuals for each group, labeling them $\tilde{e}_1$ and $\tilde{e}_2$, respectively.
3. Obtain the absolute deviations for each residual from its group median:

   $$d_{i1} = |e_{i1} - \tilde{e}_1| \qquad i = 1,\ldots,n_1 \qquad\qquad d_{i2} = |e_{i2} - \tilde{e}_2| \qquad i = 1,\ldots,n_2$$
4. Obtain the sample mean absolute deviation from the median for each group: $\bar{d}_1 = \dfrac{\sum_{i=1}^{n_1} d_{i1}}{n_1}$, $\bar{d}_2 = \dfrac{\sum_{i=1}^{n_2} d_{i2}}{n_2}$
5. Obtain the pooled variance of the absolute deviations:

   $$s^2 = \frac{\sum_{i=1}^{n_1}(d_{i1} - \bar{d}_1)^2 + \sum_{i=1}^{n_2}(d_{i2} - \bar{d}_2)^2}{n-2}$$
6. Compute the test statistic: $\mathrm{t_{BF}}^* = \dfrac{\bar{\mathrm{d}}_1 - \bar{\mathrm{d}}_2}{\mathrm{s}\sqrt{\dfrac{1}{\mathrm{n}_1} + \dfrac{1}{\mathrm{n}_2}}}$
7. Conclude that the error variance is not constant if $|\mathrm{t_{BF}}^*| \geq \mathrm{t}(1-\alpha/2;\mathrm{n}-2)$, otherwise conclude the error variance is constant.

(Notice from Step 4 to 7, it is exactly a two-sample t-test)

Look at the third part of R codes: Chapter3_Graphs_BrownTest.R

Errors are Independent (When Data Collected Over Time) (Tests for Randomness)

When data are collected over time, one common departure from independence is that error terms are positively autocorrelated. That is, the errors that are close to each other in time are similar in magnitude and sign. This can happen when learning or fatigue is occuring over time in physical processes or when long-term trends are occuring in social processes. A test that can be used to determine whether positive autocorrelation (non-independence of errors) exists is the Durbin-Watson test (see Section 12.3, we will consider it in more detail later). The test can be conducted as follows:

- H_0 : The errors are independent
- H_A : The errors are not independent (positively autocorrelated)

- $TS : D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$
- Decision Rule: (i) Reject H_0 if $D \leq d_L$ (ii) Accept H_0 if $D \geq d_U$ (iii) withhold judgment if $d_L < D < d_U$ where d_L, d_U are bounds indexed by: α, n, and p-1 (the number of predictors, which is 1 for now). These bounds are given in Table B.7, pages 674-675.

3.7 *F* Test for Lack of Fit to Test for Linear Relation between E{*Y*} & *X*

A test can be conducted to determine whether the true regression function is that which is being currently specified. For the test to be conducted, we must have the following conditions hold. The observations *Y*, conditional on their *X* level are independendent, normally distributed, and have the same variance σ^2. Further, the *X* levels in the sample must have **repeat** observations at a minimum (preferably more) of one *X* level. Repeat trials at the same level(s) of the predictor variable(s) are called *replications*. The actual observations are referred to as *replicates*.

The null and alternative hypotheses for the simple linear regression model are stated as:

$$H_0 : E\{Y \mid X\} = \beta_0 + \beta_1 X \qquad\qquad H_A : E\{Y \mid X\} = \mu_X \neq \beta_0 + \beta_1 X$$

The null hypothesis states that the mean structure is a linear relation, the alternative says that the mean structure is any structure except linear (this is not simply a test of whether β_1=0). The test (which is a special case of the general linear test) is conducted as follows:

1. Begin with *n* total observations at *c* distinct levels of *X*. There are n_j observations at the j^{th} of *X*. $n_1 + \cdots + n_c = n$
2. Let Y_{ij} be the i^{th} replicate at the j^{th} level of *X* $j = 1, \ldots, c \quad i = 1, \ldots, n_j$
3. Fit the Full model (H_A): $Y_{ij} = \mu_j + \varepsilon_{ij}$ The least squares estimate of μ_j is $\hat{\mu}_j = \overline{Y}_j$
4. Obtain the error sum of squares for the Full model, also known as the Pure Error sum of squares. $SSE(F) = SSPE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2$
5. The degrees of freedom for the Full model is df_F= *n*-*c*. This is from the fact that the j^{th} level of *X*, we have n_j-1 degrees of freedom, and they sum up to *n*-*c*. Also, we have estimated *c* parameters ($\mu_1, \ldots, \mu_c$).
6. Fit the Reduced model (H_0): $Y_{ij} = \beta_0 + \beta_1 X_j + \varepsilon_{ij}$ The least squares estimate of $\beta_0 + \beta_1 X_j$ is $\hat{Y}_j = b_0 + b_1 X_j$

7. Obtain the error sum of squares for the Reduced model, also known as the Error sum of squares. $SSE(R) = SSE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_j)^2$
8. The degrees of freedom for the Reduced model is df_R=n-2. We have estimated two parameters in this model (β_0, β_1)
9. Compute the F statistic: $F^* = \dfrac{\dfrac{SSE(R) - SSE(F)}{df_R - df_F}}{\dfrac{SSE(F)}{df_F}} = \dfrac{\dfrac{SSE - SSPE}{(n-2)-(n-c)}}{\dfrac{SSPE}{n-c}} = \dfrac{\dfrac{SSE - SSPE}{c-2}}{MSPE}$
10. Obtain the rejection region: $RR : F^* \geq F(1-\alpha; c-2, n-c)$

Note that the numerator of the F statistic is also known as the **Lack of Fit** sum of squares:

$$SSLF = SSE - SSPE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(\bar{Y}_j - \hat{Y}_j)^2 = \sum_{j=1}^{c} n_j(\bar{Y}_j - \hat{Y}_j)^2 \qquad df_{LF} = c-2$$

The degrees of freedom can be intuitively thought of as being a result of fitting a simple linear regression model of c sample means on X. Note then that the F statistic can be written as:

$$F^* = \frac{\dfrac{SSE(R) - SSE(F)}{df_R - df_F}}{\dfrac{SSE(F)}{df_F}} = \frac{\dfrac{SSE - SSPE}{(n-2)-(n-c)}}{\dfrac{SSPE}{n-c}} = \frac{\dfrac{SSE - SSPE}{c-2}}{MSPE} = \frac{\dfrac{SSLF}{c-2}}{MSPE} = \frac{MSLF}{MSPE}$$

Thus, we have partitioned the Error sum of squares for the linear regression model into Pure Error (based on deviations from individual responses to their group means) and Lack of Fit (based on deviations from group means to the fitted values from the regression model). SSE = SSPE + SSLF.

The expected mean squares for $MSPE$ and $MSLF$ are as follows:

$$E\{MSPE\} = \sigma^2 \qquad E\{MSLF\} = \sigma^2 + \frac{\sum n_j[\mu_j - (\beta_0 + \beta_1 X_j)]^2}{c-2}$$

Under the null hypothesis (relationship is linear), the second term for the lack of fit mean square is 0. Under the alternative hypothesis (relationship is not linear), the second term is positive. Thus large values of the F statistic are consistent with the alternative hypothesis.

Look at R codes: Chapter3_LackOfFit_Bank.R

3.8 Remedial Measures

If the simple linear regression model is not appropriate, there are two basic choices:

1. Abandon the simple linear regression model and develop a more appropriate one
2. Employ some transformation

Nonlinearity of Regression Function

1. Nonlinear function:

Quadratic Regression Function: $E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2$ (Places a bend in the data)

Exponential Regression Function: $E\{Y\} = \beta_0 \beta_1^X$ (Allows for multiplicative increases)

2. Transformation approach

Nonconstant Error Variance

1. Often transformations can solve this problem.
2. Another option is weighted least squares (Chapter 11).

Nonindependent Error Terms

1. Work with a model permitting correlated errors.
2. Other options include working with differenced data or allowing for previously observed *Y* values as predictors.

Nonnormality of Errors

Nonnormal errors and errors with nonconstant variances tend to occur together. Some of the transformations used to stabilize variances often normalize errors as well. The Box-Cox transformation can (but not necessarily) cure both problems.

Omission of Important Variables

When important predictors have been ommitted, they can be added in the form of a multiple linear regression model (Chapter 6).

Outliers

When an outlier has been determined to be not due to data entry or recording error and should not be removed from model due to other reasons, indicator variables may be used to classify these observations away from others (Chapter 11), or use of robust methods (Chapter 11).

3.9 Transformations

See Section 3.9 (pages 129-137) for prototype plots and transformations of *Y* and/or *X* that are useful in linearizing the relation and/or stabilizing the variance. Many times simply taking the logarithm of *Y* can solve the problems.

Box-Cox transformation (1964):

$$y' = \begin{cases} (y^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

1. Maximum likelihood method to estimate λ and other parameters $\beta_0, \beta_1, \sigma^2$ using the regression model $Y' = \beta_0 + \beta_1 X_i + \varepsilon_i$
2. Numerical search in a range of potential λ, find the λ that minimizes SSE

Comments:

1. Residual plots and other analyses described earlier should be employed to ascertain that the simple linear regression model is appropriate for the transformed data
2. If the Box-Cox procedure leads to λ=1, then no transformation of Y is needed.

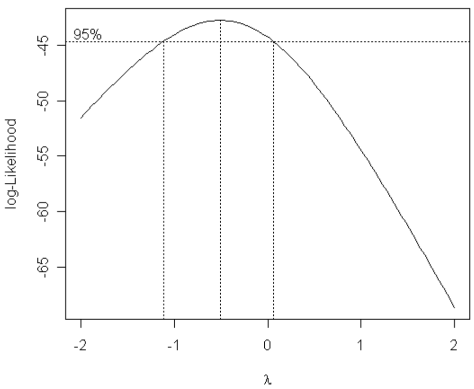
An Example: Consider the plasma level example (pg 133).
Data on age (X) and plasma level of a polyamine (Y) for a portion of the 25 healthy children are presented in CH03TA08.txt. Check for the log-transformation.

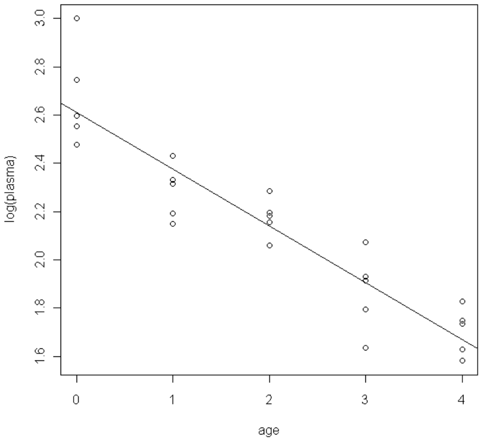
Look at the R codes: Chapter3_Plasma_BoxCoxTransformation.R

R codes:

```
Data <- read.table(file="C:/…/DataSet/CH03TA08.txt", header=FALSE)
# age plasma lplasma
age <- Data[,1]
plasma <- Data[,2]

library(MASS) #call the library MASS (Modern Applied Statistics with S)
boxcox(plasma~age)
```

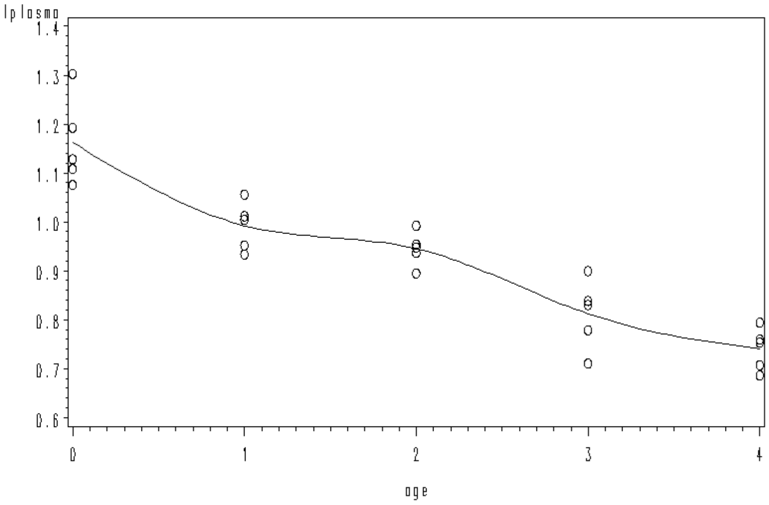




3.10 Exploration of Shape of Regression Function

Smoothing methods such as cubic smoothing spline, loess, lowess, can be used not only for exploring regression relationship but also for confirming the nature of regression function.

For example, we can replace i=SMxx in the symbol statement in the previous SAS example. It specifies that a smooth line be fit to noisy data using a spline routine. The points on the plot do not necessarily fall on the line. Specifying I=SMxx results in fitting a cubic spline that minimizes a linear combination of the sum of squares of the residuals of fit and the integral of the square of the second derivative (Reinsch, 1967). The value xx can range from 01 to 99 and determines the relative importance of the two components: the larger the value, the smoother the fitted curve.



For R, the smoothing functions include: smooth.spline, loess, lowess. To find out more, type the following command in R:

```
> help(smooth.spline)
```