Chapter 3: Diagnostics and Remedial Measures

Introduction:

- Recall that a residual for the $i$th observation is defined as
  \[ e_i = Y_i - \hat{Y}_i \]

- Recall that in the model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, we assume that $\epsilon_i \sim N(0, \sigma^2)$ where $\epsilon$ is unobservable.

- Residuals may be regarded as observed error.

- If the model is appropriate, the residuals should reflect the properties assumed by the model error.

We will study the following departures from the model through the residuals.

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.

Diagnostics for Residuals

We will use residual plots as an informal method of analysis. Below is a summary of which plots are used to assess specific departures from the model assumptions.

<table>
<thead>
<tr>
<th>Plot</th>
<th>Departure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residuals Vs Predictor Variables</td>
<td>1, 2, 4</td>
</tr>
<tr>
<td>Absolute (or Squared) Residuals Vs Predictor Variables</td>
<td>2</td>
</tr>
<tr>
<td>Residuals Vs Fitted values</td>
<td>1, 2, 4</td>
</tr>
<tr>
<td>Residuals Vs Time (or other sequence)</td>
<td>3</td>
</tr>
<tr>
<td>Residuals Vs Omitted Predictor Variables</td>
<td>6</td>
</tr>
<tr>
<td>Box Plot of Residuals</td>
<td>4, 5</td>
</tr>
<tr>
<td>Normal probability plot of residuals</td>
<td>5</td>
</tr>
</tbody>
</table>

Non-linearity of the regression function

- Recall that we want to explain the relationship $Y = f(X)$ where - for our class - $f(X)$ is assumed to be a linear function.
• We will consider two plots for this analysis.
  – Residuals Vs Predictors (i.e. $e_i$ Vs $X_i$)
  – Residuals Vs Fitted Values (i.e. $e_i$ Vs $\hat{Y}_i$)

• What are we asking of this diagnostic? Is a linear regression function appropriate?

• What are we looking for? **Systematic patterns**
  
  Examples
Notes:

- For the simple linear regression case, the plots $e_i$ Vs $X_i$ and $e_i$ Vs $\hat{Y}_i$ reveal identical information.
- For multiple regression, separate plots for $e_i$ Vs $(\hat{Y}_i, X_{1i}, X_{2i}, \ldots, X_{ki})$ will be useful.

**Non-constancy of error variances**

- There are two basic types of plots with some variations.
- We will consider the following plots.
  - Residuals Vs Predictors (or Fitted Values)
  - Absolute Residuals Vs Predictor (or Fitted Values)
  - Squared Residuals Vs Predictor (Or Fitted Values)
- The magnitude of the residuals is more important in this diagnostic rather than the sign.
- What are we looking for? Systematic patterns showing that variation in the residuals changes over the range of the predictor(s) (or fitted values).

Examples
**Presence of Outliers**

Outlier (defined) - an extreme observation.

Graphical tools to identify outliers:
- $e_i$ Vs $X_i$ (or the $\hat{Y}_i$’s).
- box plot
- stem and leaf plots
- dot plot

A box plot provides a graphic of the 5 - number summary of data.

A stem and leaf plot and dot plot can be viewed as a histogram-type graphic.
Consider the ordered test scores on a calculus test

42, 43, 47, 56, 65, 72, 75, 81, 82, 87, 88, 90, 91
Using Standardized Residuals

• We can also use the standardized residuals for each of the previous plots.

• The standardized residuals are calculated as

\[ e_i^* = \frac{e_i}{\sqrt{MSE}} \]

• Standardizing, in this case, rescales the variance of \( e_i \) to 1.

• Recall the relationship of a \( z \)-score to the normal distribution where

\[ z = \frac{x - \mu}{\sigma} \sim N(0, 1) \]

• Also, recall the empirical rule which states that if the data are bell-shaped then
  - 68% of the data lie between ±1 standard deviation.
  - 95% of the data lie between ±2 standard deviations.
  - 99.5% of the data lie between ±3 standard deviations.

• By standardizing our residuals the standard deviation now is also 1.

• Consider the plot below of the standardized residuals against the predictor \( X \).
Notes on Outliers:

- Outliers create problems in regression.
- The reason for this is that the Method of Least Squares provides estimates that are influenced by outliers.
- For small sample sizes, an outlier will influence regression estimates more than for large sample sizes.

Sources of Outliers:

- Mistake or error in recording.
- Miscalculation.
- Malfunctioning equipment.
- Omission of a predictor variable in the model.

As a general rule, if there is direct evidence supporting any of the first three sources then one may discard the observation.
Non-independence of Error Terms

If our data has been collected over a time sequence (or geographical region) then the residuals should be reviewed for correlation.

Below are three prototypes of graphs with correlated residuals and one showing independent residuals.

- Interpretation of residual plots can be tricky.
- The specific plot and the shape of the plot are critical to correct interpretation.
- Below is a prototype dependence plot versus one showing a poorly fitting model.
Non-normality of Error Terms

- Most difficult of the departures to detect.
- Without large sample sizes detection can be tricky.
- We can assess the symmetry behavior of the residuals through box plots, histograms, dot plots, and stem and leaf plots.
- These are good to assess gross departures but require large sample sizes.
- We can use a normal probability plot to assess the frequency behavior of the residuals.

Normal probability plot of residuals (rationale) - the sample percentiles of the ordered residuals should approximate the percentiles of a normal distribution.

- We can use either the raw residuals or the standardized residuals.
- Begin by ordering the residuals.
- Calculate the expected value of the residual under the assumption of normality (which will use the rank).
- Plot Observed residuals against Expected residuals.
- Under normality, we expect a straight line at a 45° angle.

Examples
Omission of Important Predictor Variables

- It could be the case that an important predictor variable has been excluded from the model.
- We can plot residuals against omitted predictors to assess this scenario.
- What are we looking for? Systematic trends in the variation of the residuals with the level of the omitted predictors.
- We could improve the model fit by including important predictors.
- Example - Salary increases as a function of an individual’s height.

Objective Tests for Assessing Validity of Assumptions

- Although graphical analysis is subjective, in many cases it suffices for examining the aptness of the model.
- A hypothesis test, however, puts specific questions to the test.
- We will consider only two objective tests:
  - The Bruesch-Pagan Test for Constancy of Error Variance.
  - The Lack of Fit Test to assess the appropriateness of a linear function.

Breusch-Pagan Test

- If $\sigma_i^2$ is not constant, we may expect variation for different levels of $X$ (our predictor).
- Consider the model
  \[ \ln \sigma_i^2 = \gamma_0 + \gamma_1 X_i \]
- This model takes into account that $\sigma_i^2$ can either increase or decrease with $X_i$.
- If $\gamma_1 = 0$ then we could conclude that $\sigma_i^2$ does not vary with $X_i$.
- Consider the following hypothesis:
  \[ H_0 : \gamma_1 = 0 \quad Vs \quad H_A : \gamma_1 \neq 0 \]
  as a hypothesis test for nonconstancy of variance.
- That is
  \[ H_0 : \sigma_i^2 \text{ is constant for all } \epsilon_i \]
  \[ H_A : \sigma_i^2 \text{ is not constant} \]
- The test statistic $\chi^2_{BP}$:
Let SSR* be the Regression Sum of Squares from regressing $e^2_i$ on $X_i$.
Let SSE be the usual Sum of Squares Error computed from regressing $Y_i$ on $X_i$.
Let $n$ denote the sample size.
Our test statistic is given by:

$$\chi^2_{BP} = \frac{SSR^*/2}{(SSE/n)^2} \sim \chi^2(df = 1)$$

Reject the null hypothesis for $\chi^2_{BP} > \chi^2_{(1-\alpha)}$

*F-test for Lack of Fit*

- We have assumed that our response, $Y$, is related linearly to our predictor $X$.
- Recall our model $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$.
- Thus, $E[Y] = \beta_0 + \beta_1 X$.
- Our hypothesis to test for lack of fit is given by:

$$H_0 : E[Y] = \beta_0 + \beta_1 X \quad \text{Vs.} \quad H_A : E[Y] \neq \beta_0 + \beta_1 X$$

- This test requires repeat observations at one or more levels of $X$.
- These replicates can either happen by design or by chance.
- The test statistic requires defining a Reduced and Full Model.

**Reduced Model for F-test for Lack of Fit:**

- Let $c$ be the number of different levels of $X$.
- Let $n$ be the total number of observations
- Under the null hypothesis, the reduced model is given by

$$Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, c$.
- We will define SSE(R) as the Sum of Squared Error for the Reduced Model.

**Full Model for F-test for Lack of Fit:**

- Under the full model we do not assume a specific form.
• The full model is given by
\[ Y_{ij} = \mu_j + \epsilon_{ij} \]
where \( i = 1, \ldots, n \) and \( j = 1, \ldots, c \) and \( \epsilon_{ij} \sim N(0, \sigma^2) \).

• From our model we have that \( E[Y] = \mu_j \).

• An estimate of the \( \mu_j \) is given by \( \bar{Y}_j \).

• We will define \( \text{SSE}(F) \) as the sum of squared error for the full model.

• \( \text{SSE}(F) \) is also called Pure Error Sum of Squared (SSPE) and is given by
\[
\text{SSE}(F) = \sum_{i=1}^{n} \sum_{j=1}^{c} (Y_{ij} - \bar{Y}_j)^2
\]

Test Statistic for \( F \)-test for Lack of Fit:

• Our test statistic and distribution under the null hypothesis is given by
\[
F^* = \frac{\frac{\text{SSE}(R) - \text{SSE}(F)}{c-2}}{\frac{\text{SSE}(F)}{n-c}} \sim F(d f_1 = c-2; \ d f_2 = n-c)
\]

• We reject \( H_0 \) if \( F^* > F_{1-\alpha; \ d f_1 = c-2, \ d f_2 = n-c} \)

• If we reject \( H_0 \) then we conclude a linear model is not adequate to describe the data.

• Note the use of \texttt{PROC ANOVA} in SAS to calculate SSPE.