Diagnostics and Remedial Measures I

Notes from “Linear Statistical Models” by Kutner et al

Devere Anthony Weaver

1 Overview

When a regression model, such as the simple linear regression model, is selected for an application, one can usually not be certain in advance that the model is appropriate for that application. Any of the features of the model may not be appropriate for the particular data at hand, thus we need to examine the aptness of the model for the data before further analysis based on that model is undertaken.

As such, there are a few different ways to help determine model aptness. There are some simple informal graphical methods as well as formal statistical tests.

1.1 Diagnostics for Independent Variable

Some useful graphic diagnostics for the independent variable contains a simple box plot, time plots, steam-and-leaf plot, dot plot, etc.

2 Residuals

Direct diagnostics for the dependent variable Y are not too useful in regression analysis because the values of the observations on the dependent variable are a function of the level of the independent variable.

Instead, diagnostics for the dependent variable are usually carried out through an examination of the residuals. If the model is appropriate for the data at hand, the observed residuals should then reflect the properties assumed for the model's residuals.

2.1 Properties of Residuals

The mean of the n residuals e_i for the simple linear regression model is

$$\bar{e} = \frac{\sum e_i}{n} = 0 \tag{2.1}$$

where $\bar{e}$ denotes the mean of the residuals. The variance of the n residuals e_i is given by the MSE.

Standardized residuals are used at times in residual analysis.

2.2 Departures from Model to Be Studied by Residuals

The following are six important types of departures from the linear regression model with normal errors:

1. The regression function is not linear.
2. The error terms don't 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 aren't normally distributed.
6. One or several important independent variables have been omitted from the model.

---

3 Diagnostics for Residuals

We can use informal diagnostic plots of residuals to provide information on whether any of the six types of departures from the simple linear regression model are present. The following plots of residuals will be utilized for this purpose:

1. Plot of residuals against independent variable
2. Plot of residuals against fitted values
3. Plot of residuals against time
4. Plot of residuals against omitted independent variable
5. Box plot of residuals
6. Normal probability plot of residuals

---

4 Nonlinearity of Regression Function

Whether a linear regression function is appropriate for the data being analyzed can be studied from a *residual plot against the independent variable* or from a *residual plot against the fitted values*, and also from a *scatter plot*. However, a scatter plot is not always as effective as a residual plot. One thing a scatter plot is good for is potentially checking the functional form of the relationship between the variables.

A residual plot has some advantages over a scatter plot including it can easily be used for examining other facets of the aptness of the model, scaling isn't as much of an issue compared to a scatter plot, and they can clearly show any systematic pattern in the deviations around the fitted regression line.

For a residual plot against the independent variable, the residuals should tend to fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive and negative.

4.1 Nonconstancy of Error Variance

Plots of the residuals against the independent variable or against the fitted values are not only helpful to study whether the variance of the error terms is constant.

One of the telltale signs of nonconstant variance in the error terms is the fanning shape when the residuals are plotted against X.

4.2 Presence of Outliers

Residual outliers can be identified from *residual plots against* X *or* $\hat{Y}$, as well as from box plots, steam-and-leaf plots, and dot plots. In a standardized residual plot, outliers are points that lie far beyond the scatter of the remaining residuals.

Outliers can be difficult to deal with, as such, a safe rule frequently suggested is to discard an outlier only if there is direct evidence that it represents an error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstance.

4.3 Overview of Tests Involving Residuals

So, there are many different tests and they are covered in more depth in a later chapter that goes more into detail about all the different available tests.

The only one that's covered here is the F test for lack of fit. You can use the output of the software package to see the result of this test.

Note: This will do it for the coverage of the notes. For more detail, reference the text or other resources, this is just for quick reference.

---

5 Transformations

Simple transformations of either the dependent or the independent variable, or both, are often sufficient to make the simple linear regression model appropriate for the transformed data.

Again, the coverage here can be technical and isn't worth copying here to this document. However, Kuter et al wrote a great section on this and how to determine the potential transformation based on the prototype regression pattern. Use this text as a reference when dealing with potentially transforming variables.