

Chapter 5

Residual Analysis

In the regression analysis, we always assume that the error term satisfies:

(i) normally distributed with mean 0, (ii) the variance is constant, (iii) errors are independent.

This chapter provides both graphical tools and statistical tests that will aid in checking the validity of these assumptions. It also takes up a number of diagnostics for detecting improper functional form for a predictor variable, outliers, and influential observations.

Throughout this chapter, we assume that there are k predictor variables.

5.1 Plotting Residuals and Detecting Lack of Fit

Consider the model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

and define the regression residual

$$\hat{\varepsilon} = y - \hat{y} = y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k)$$

Detecting model lack of fit with residuals:

- Plot the residuals, $\hat{\varepsilon}$, on the vertical axis against each of the predictor variables, x_1, x_2, \dots, x_k on the horizontal axis
- Plot the residuals, $\hat{\varepsilon}$, on the vertical axis against the predicted value, \hat{y} , on the horizontal axis
- In each plot, look for trends, dramatic changes in variability, and/or more 5% of residuals that lie outside $2s$ of 0. Any of these patterns indicates a problem with model fit.

Partial regression residuals:

The partial regression residuals for the j th predictor variable x_j is defined by

$$\hat{\varepsilon}^*(j) = y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{j-1} x_{j-1} + \hat{\beta}_{j+1} x_{j+1} + \hat{\beta}_k x_k) = \hat{\varepsilon} + \hat{\beta}_j x_j.$$

- Plot $\hat{\varepsilon}^*(j)$ against x_j . The points should be scattered around a line with slope equal to $\hat{\beta}_j$. Unusual deviations or patterns around this line indicate lack of fit for the variable x_j .

Example 5.1 Refer to [Example 8.3, p.372] and [Case Study 13, p. 652].

5.2 Detecting Unequal Variances

Recall that one of the assumptions necessary for the validity of regression inferences is that the error term ε have constant variance for all levels of the predictor variables.

homoscedastic: equal variances

heteroscedastic: unequal variances

Residual plot against \hat{y} :

- Plot the residuals, $\hat{\varepsilon}$, on the vertical axis against the predicted value, \hat{y} , on the horizontal axis

Example 5.2 Consider the data set given in [Table 8.4, p.382]. The response variable is *salary* y , the predictor variable is *years of experience*.

Test for unequal variances:

Divide the data into **two subsamples**. Assume the sample 1 has size n_1 and sample 2 has size n_2 .

$H_0 : \sigma_1^2 = \sigma_2^2$ (assumption of equal variances satisfied)

$H_a : \sigma_1^2 \neq \sigma_2^2$ (assumption of equal variances violated)

where

σ_1^2 = variance of the random error for subpopulation 1,

σ_2^2 = variance of the random error for subpopulation 2.

Test statistic:

$$F = \frac{\text{Larger MSE}}{\text{Smaller MSE}}$$

Under H_0 , F has F-distribution with v_1 and v_2 degrees of freedom, where v_1 = degree of freedom associated with the larger MSE, v_2 = degree of freedom associated with the smaller MSE.

5.3 Checking the Normality Assumption

Recall that all the inferential procedures associated with a regression analysis are based on the assumptions that the random error is **normally distributed** with mean 0 and variance σ^2 .

The **normal probability plot** (normal quantile plot) is a commonly used graphical technique for checking the assumption of normality.

Constructing a normal probability plot for regression residuals:

- List the residuals in ascending order, say, $e_{(1)} \leq e_{(2)} \leq \dots \leq e_{(n)}$.
- Let $p_i = \frac{i - .375}{n + .25}$ and choose z_i so that $P(N(0, 1) \leq z_i) = p_i$.
- Plot $e_{(i)}$ against z_i
- If the normality assumption is satisfied, then the plot should show **a linear trend**.

5.4 Detecting Outliers and Identifying Influential Observations

Identifying Outlying Observations

- Use the **standardized residual**:
 $z_i = e_i/s$, where $e_i = y_i - \hat{y}_i$.
If $|z_i| > 3$, then the observation is considered to be an **outlier**

- Use Studentized Deleted Residual:

To calculate the Studentized deleted residual for the i th observation, we first remove the i th observation from the data set and then calculate the regression. Define the [deleted residual](#) by

$$d_i = y_i - \hat{y}_{i(i)}$$

Then

$$s_{d_i}^2 = \frac{\text{MSE}_{(i)}}{1 - h_{ii}}$$

Studentized deleted residual:

$$\begin{aligned} t_i &= \frac{d_i}{s_{d_i}} \\ &= e_i \left(\frac{n - 2 - k}{\text{SSE}(1 - h_{ii}) - e_i^2} \right)^{1/2} \end{aligned}$$

where $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = (h_{ij}, 1 \leq i, j \leq n)$.

If the i th observation is not an outlier, then t_i has a t-distribution with $n - 2 - k$ degrees of freedom.

- Use Hat Matrix Leverage:

The diagonal element h_{ii} of the hat matrix is a useful indicator whether or not the i th observation is outlying with respect to its x values.

- $0 \leq h_{ii} \leq 1$, $\sum h_{ii} = k + 1$.
- h_{ii} is a measure of the distance between the x values of the i th observation and the mean of the x values for all n cases.
- A large value h_{ii} indicates that the i th case is distant from the center of x .
- h_{ii} is considered [large](#) if $h_{ii} > 2p/n$.

Identifying Influential Observations

We shall consider an observation to be [influential](#) if its exclusion causes major changes in the fitted regression function.

- Influence on single fitted value - DFFITS:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

Guideline:

A case is considered **influential** if $|DFFITS| > 1$ (small data set) or $|DFFITS| > 2\sqrt{(k+1)/n}$ for large data set.

- Influence on all fitted values - Cook's distance:

Cook's distance measure considers the influence of the i th case on all n fitted values.

$$\begin{aligned} D_i &= \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{(k+1) \text{MSE}} \\ &= \frac{e_i^2}{(k+1) \text{MSE}} \frac{h_{ii}}{(1 - h_{ii})^2} \end{aligned}$$

- D_i becomes large with either a poor fit (e_i is large) or high leverage (h_{ii} close to 1) or both
- $D_i \approx F(k+1, n-k-1)$
- If $P(F_{k+1, n-k-1} \leq D_i) > .5$, then the i th case has **major influence** on the fit of the regression function.

5.5 Detecting Residual Correlation: The Durbin-Watson Test

Time series: data are observed over time.

Main features of a time series:

- The value at time t is **correlated** with the value at time $(t+1)$.
- seasonal variation
- trend

If a time series is used as the predictor variable in a regression analysis, then the random errors are **correlated**.

The Durbin-Watson d statistic:

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

Properties:

- Range of d : $0 \leq d \leq 4$
- If residuals are **uncorrelated**, $d \approx 2$
- If residuals are **positively correlated**, $d < 2$, and if the correlation is very strong, $d \approx 0$
- If residuals are **negatively correlated**, $d > 2$, and if the correlation is very strong, $d \approx 4$

Example 5.3 *The table below is part of the city average price of gasoline over the period Sept. 1981 to March 2004.*

09/15/1981	1.471
10/15/1981	1.470
11/15/1981	1.470
12/15/1981	1.468
⋮	⋮

5.6 Detecting Multicollinearity

In multiple regression analysis, the nature and significance of the relations between the predictor variables and the response variable are often of particular interest.

Ideal situation: predictor variables are uncorrelated.

Multicollinearity:

Two or more of the predictor variables used in regression are **moderately or highly** correlated.
Example of perfectly correlated predictor variables:

x_1	x_2	y
2	6	23
8	9	83
6	8	63
10	10	103

Solution:

Two key implications of this example are:

- The perfect relation between X_1 and X_2 did not prohibit our ability to obtain a good fit to the data.
- Since many different regression functions provide the same good fit, we cannot interpret any one set of regression coefficients as reflecting the effects of the different predictor variables.

Effects of multicollinearity:

- Adding or deleting a predictor variable changes the regression coefficients.
- The estimated standard deviations of the regression coefficients become large when the predictor variables in the regression model are highly correlated with each other.
- The estimated regression coefficients individually may not be statistically significant even though a definite statistical relation exists between the response variable and the set of predictor variables.

Variance Inflation Factor:

The variance inflation factor (VIF) for the i th regression coefficient is defined by

$$\text{VIF}_i = \frac{1}{1 - R_i^2},$$

where R_i^2 is the [coefficient of multiple determination](#) of the regression produced by regressing the variable x_i against the [other predictor variables](#).

Diagnostic Uses:

The largest VIF value among all X variables is used as an indicator of the [severity of multicollinearity](#). A maximum VIF value in excess of **10** is taken as an indication that multicollinearity may be unduly influencing the least squares estimates.

Example 5.4 Refer to *[Example 7.3, p.349]*.