Regression Diagnostic and Residual Analysis

A regression analysis is not complete until one is convinced the model is an adequate representation of the data. An examination and an analysis of the residuals are crucial components of the determination of the model adequacy.

How to Analyze?

1) Normal probability plot of residuals  
2) Histogram of residuals  
3) Residuals vs Fits  
4) Plot of residuals over time if the data are chronological  
5) AC function graph of residuals, if regression model is used with time series data.  
6) Computing the measure of influence of the ith data point (leverage)

\[
h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}
\]

is the leverage computation formula for simple linear regression. 0<\(h_{ii}\)<1.

If \(h_{ii} \geq 3(k+1)/n\), \(Y_i\) has a large influence on the model parameters’ estimated values.

7) One can also compute the standardized residuals. For simple linear regression, standardized value for any \(e_i\) is computed by the ratio \(Z_{ei} = (e_i/s_{ei})\), where \(S_{ei} = s (1-h_{ii})^{1/2}\), and \(s=(MSE)^{1/2}\).

If \(|Z_{ei}| > 2\), then the corresponding \(Y_i\) is an outlier.

8) A linear relation between two or more independent variables is called multicollinearity. The strength of the multicollinearity is measured by the variance inflation factor (VIF).

\[
VIF_j = \frac{1}{R_j^2}, \text{ j=1,2,...,k}
\]

Here \(R_j^2\) is the coefficient of determination from the regression of the jth independent variable on the remaining (k-1) independent variables.

If the jth independent variable is not related to the remaining X’s, \(R_j^2 = 0\) and VIF=1. If there is a relationship, then VIF>1. If VIF is near 1 then multicollinearity is not a problem.

Regression with Time Series Data

AC exists when successive observations over time are related to one another. (Serial Correlation)

A common kind of Ac is one in which the error term in the current time period is directly related to the error term in the previous time period. In this case, the simple linear regression model takes the form:
\[ Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \]

With \( \varepsilon_t = \rho \varepsilon_{t-1} + \nu_t \) here \( \rho \) lag 1 AC coefficient, and

\[ \nu_t \sim N(0, \sigma^2_{\nu}) \]. If \( \rho = 0 \) then there is no serial AC.

If regression models are used with AC’ed time series data, it is especially important to examine the residuals.

**Durbin-Watson Test for Serial Correlation**

Consider, \( \varepsilon_t = \rho \varepsilon_{t-1} + \nu_t \). The hypotheses to be tested are: \( H_0 : \rho = 0 \) vs \( H_a : \rho > 0 \)

Since the business and economic time series tend to have positive AC.

\[
    DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = 2(1 - r_1(e)), \text{ and we an see that } 0 < DW < 4.
\]

If \( DW \equiv 2 \) then there is no serial correlation.

The **DW test** is used to determine whether positive AC is present,

- If \( DW > U \), conclude \( H_0 : \rho = 0 \).
- If \( DW < L \), conclude \( H_a : \rho > 0 \).
- If \( L \leq DW \leq U \), the test is inconclusive. In this case check to see if \( \frac{2}{\sqrt{n}} < r_1(e) < \frac{2}{\sqrt{n}} \), if so then serial AC can be ignored.

**Solution to AC Problems:**

1) One or more key variables may be omitted. Include them in the model.
2) Differencing may provide a reasonable solution to the AC problem.
3) Transformation of the original data, for instance work with \( \log(Y) \) rather than with \( Y \) itself.
4) Autoregressive models may be another form of solution to the AC problem.

**Autoregressive Models**

\[ Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t \]

An Autoregressive model expresses a forecast as a function of previous values of the time series.
The Problem of Heteroscedasticity (non-constant variability)

A simple transformation may be a solution to the problem. For instance, using a loglinear model may be a solution to heteroscedasticity.

Using Regression to Forecast Seasonal Data

\[ Y_t = \beta_0 + \beta_1 t + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_4 + \epsilon_t \]  Where \( S_i, i=2,3,4 \) are the dummy variables.

If there exists seasonality but not a trend, then \( \beta_1 = 0 \).