Autocorrelation: form of dependence that is evident from the residual plots, which is autocorrelation. This happens when ordered observations are dependent on prior observations.

One of the most common instances of this correlation structure is through data that is collected over time, also referred to as time series.

To diagnose autocorrelation in data, you can

1. Visually inspect the plot of residuals by time (see a pattern over time…usually cyclical)
2. Conduct a Durbin-Watson test

In using the Durbin-Watson test, the hypotheses are:  
H0:0: No residual correlation  
HA:: Residual Correlation

e� represents the error terms, or residuals. The statistic d ranges from 0 to 4, with a value of 2 indicating no residual correlation. Values of d smaller than 2 may indicate a positive autocorrelation and a value of d larger than 2 may indicate a negative autocorrelation.

testing for positive autocorrelation would be the preferred test.\

Durbin-Watson test is in the lmtest package of R. If you want to test for positive autocorrelation, you will need to specify the alternative of “greater”

A white screen with black text

Description automatically generated

How to identify potential outliers and influential points (what measures are used for each)

*Outliers:* Outliers tend to be those values that do not follow the trend of the data and are generally found by large deviations in the y direction.

outliers are found by using residuals

* rstandard
* rstudent

we first need to “standardize” them, or divide by their standard errors. In this sense, we can think of these “standardized” residuals as an approximate z-score. Therefore, we look for residuals greater in magnitude than 3 as potential outlier( large sample) >2 for small

*Influential observations:* Influential points are usually those points on the edges of the x-values and can greatly impact the slopes in the regression equation

Influential points can be discovered by Cook’s D, dffits, dfbetas or Hat values.

* Dffits: measures impact of ith observation on predicted value 2\*sqrt(p/n)
* Dfbetas: change in jth parameter estimate with deletion of ith observation. 2\*sqrt(1/n), one DFbetea per parameter per observation, helpful to know on which parameter the influence most lies.
* cooks.distance: Cook’s D, measures the difference in the regression estimates when the ith��ℎ observation is left out; D>4/(n-p-1)
* hatvalues 2(p)/n

How to handle influential obs?

Recheck the data to ensure that no transcription or data entry errors occurred.

2. If the data is valid, one possible explanation is that the model is not adequate. ❑ A model with higher-order terms, such as polynomials and interactions between the variables, might be necessary to fit the data well. ❑ Nonlinear model

3. Determine the robustness of the inference by running the analysis both with and without the influential observations.

4. Robust Regression (Covered Later in Program)

5. Weighted Least Squares (WLS

Multicollinearity (what it is and how to identify it)

occurs when one or more predictor variables are linearly related to each other and will create issues with the regression.

The parameter estimates will not be stable and the standard errors will be inflated (making it more difficult to find significant explantory variables).

The two most common ways to identify multicollinearity is by looking at the correlation among the predictor variables and calculating the variance inflation factor.

To deal with multicollinearity, we can do either of the following:

1. Remove one or more variables that are co-linearly related to another variable(s).
2. Create new transformed variables (take linear combinations of variables; create ratio of variables, etc).

Looking at correlation matrix of predictors

❑ One of the most commonly used measures is the variance inflation factor (VIF). will take each explanatory variable and model it as a linear regression of the remaining explanatory variables, calcua;te Rsquare of each (R1sq…)

❑ Values of VIF greater than 10 indicate potential collinearity (if using GVIF, then use a cutoff of 5) Collinearity Diagnostics 𝑉𝐼𝐹𝑖 = 1/ 1 − 𝑅𝑖