Regression Diagnostics Procedures

ASSUMPTIONS UNDERLYING REGRESSION/CORRELATION

NORMALITY OF VARIANCE IN Y FOR EACH VALUE OF X

For any fixed value of the independent variable X, the distribution of the dependent variable Y is normal.

NORMALITY OF VARIANCE FOR THE ERROR TERM

The error term is normally distributed. (Many authors argue that this is more important than normality in the distribution of Y).

THE INDEPENDENT VARIABLE IS UNCORRELATED WITH THE ERROR TERM

ASSUMPTIONS UNDERLYING REGRESSION/CORRELATION (Continued)

HOMOSCEDASTICITY

It is assumed that there is equal variances for Y, for each fixed value of X.

LINEARITY

The relationship between X and Y is linear.

INDEPENDENCE

The Y's are statistically independent of each other.

Residual

The difference between observed and predicted values of the response variable, $Y - \hat{Y}$, is called a ***residual***.

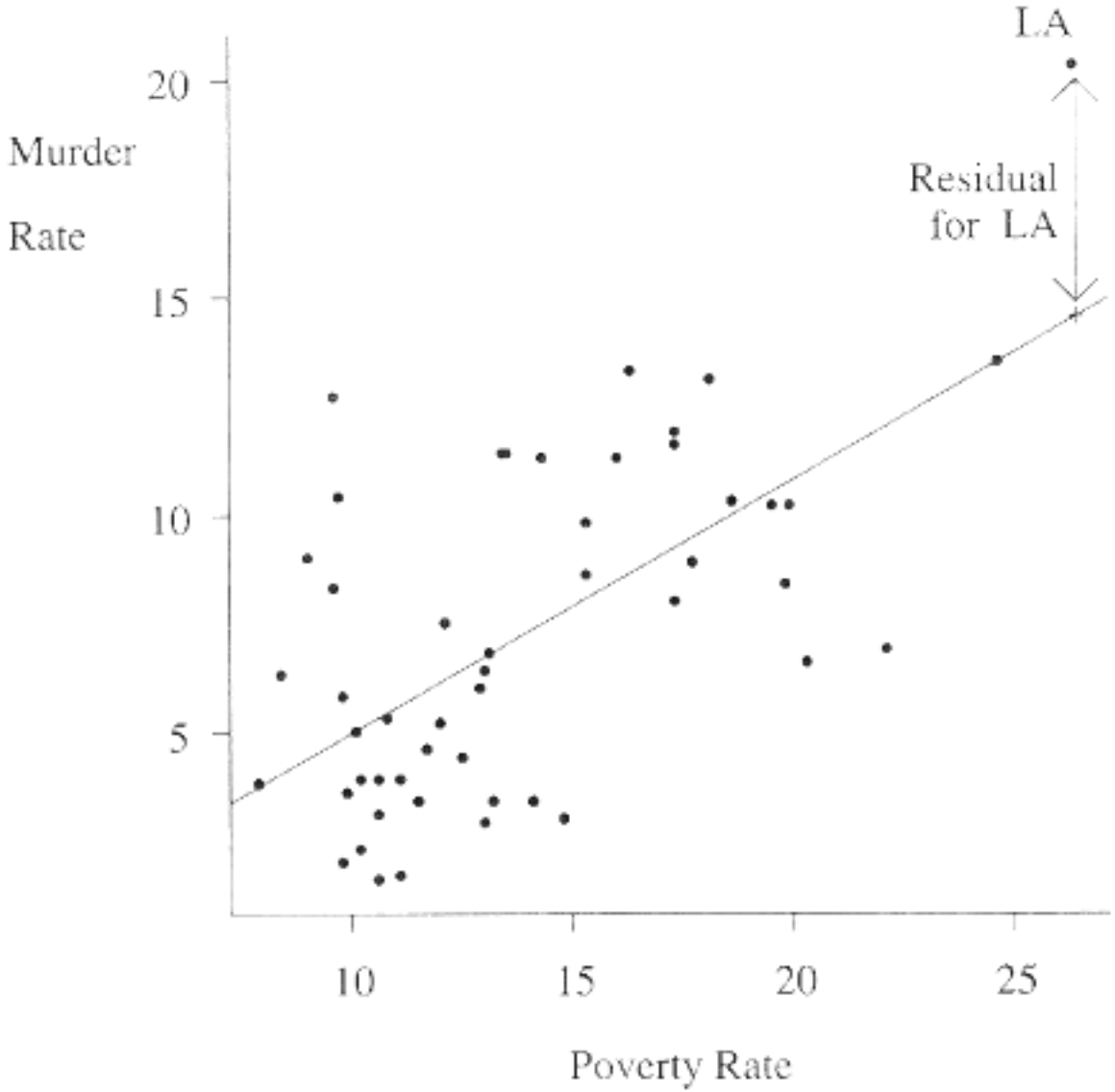

Figure 9.7 Prediction Equation and Residuals

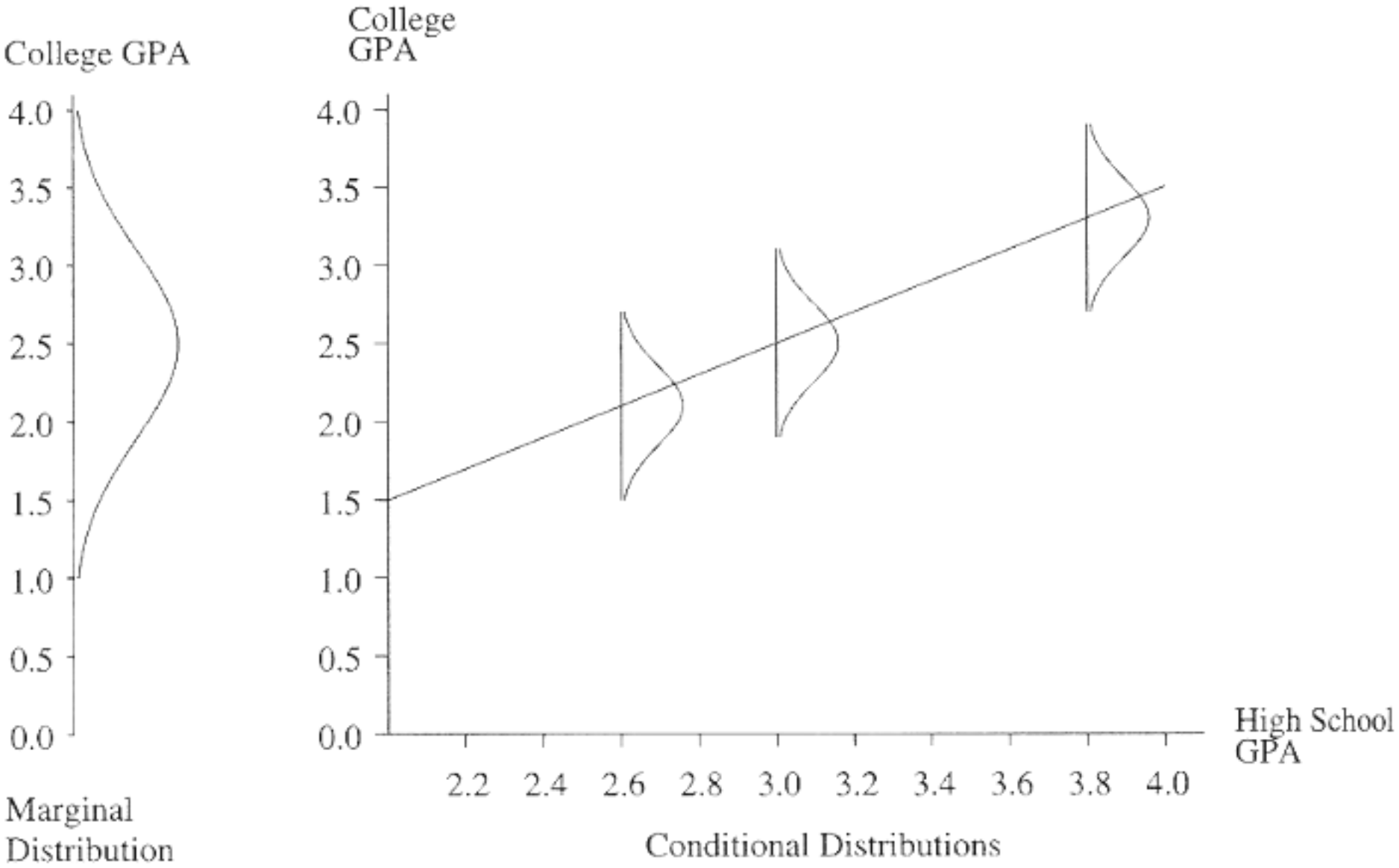

Figure 9.9 Marginal and Conditional Distributions

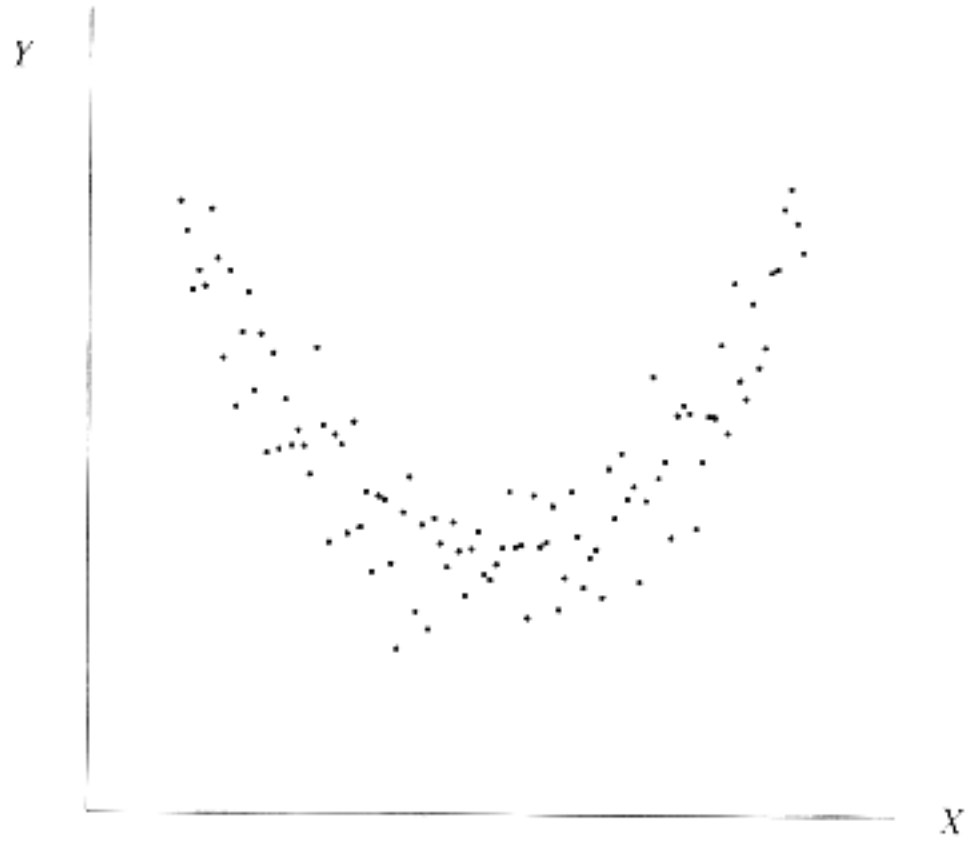

Figure 9.5 A Nonlinear Relationship, for Which It Is Inappropriate to Use a Straight Line Regression Model

Prediction Equation

When the scatter diagram suggests that the linear model $Y = \alpha + \beta X$ is realistic, we estimate this unknown line. The notation

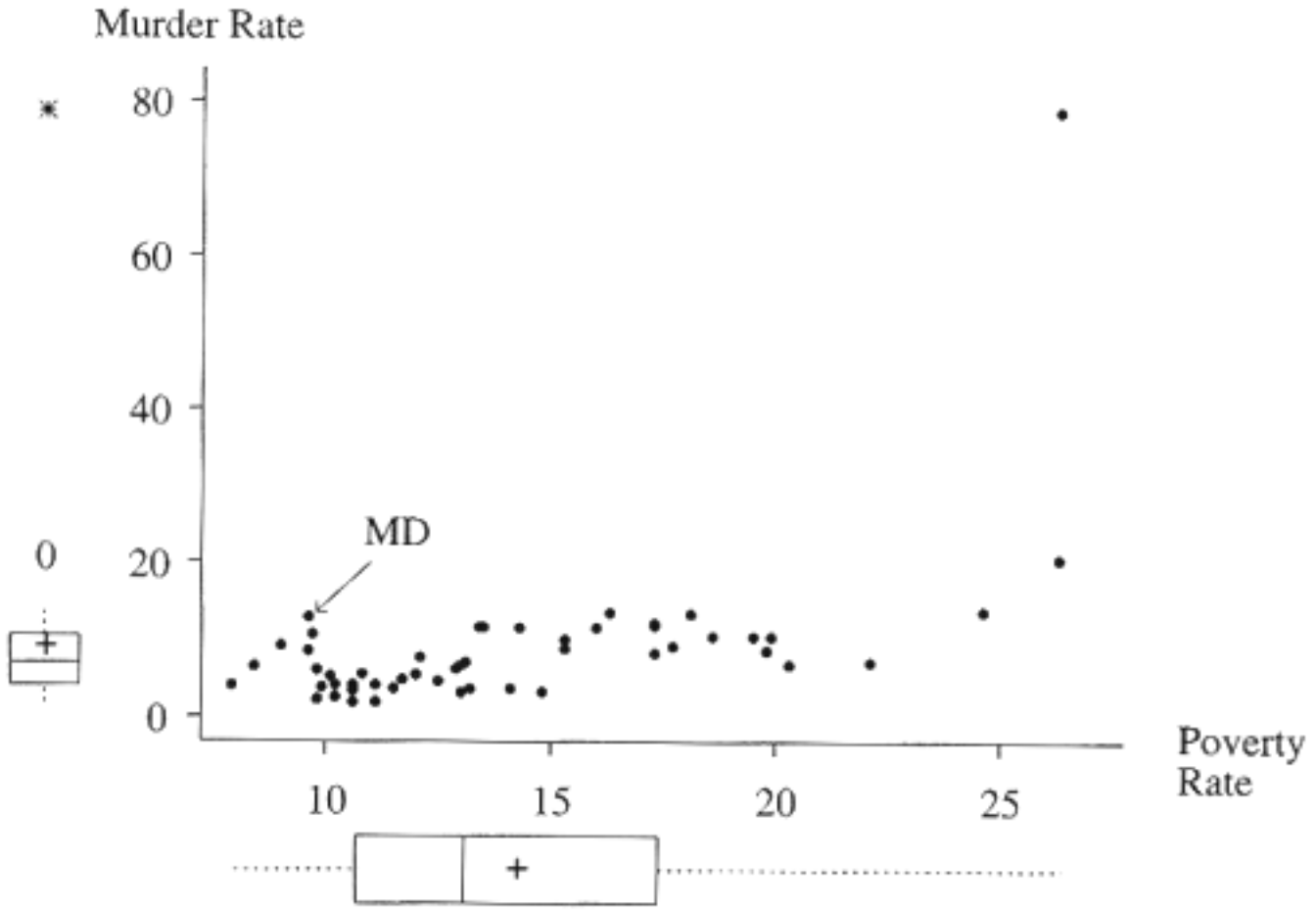

Figure 9.4 Scatter Diagram for Y = Murder Rate and X = Percentage of Residents Below the Poverty Level, for 50 States and D.C.

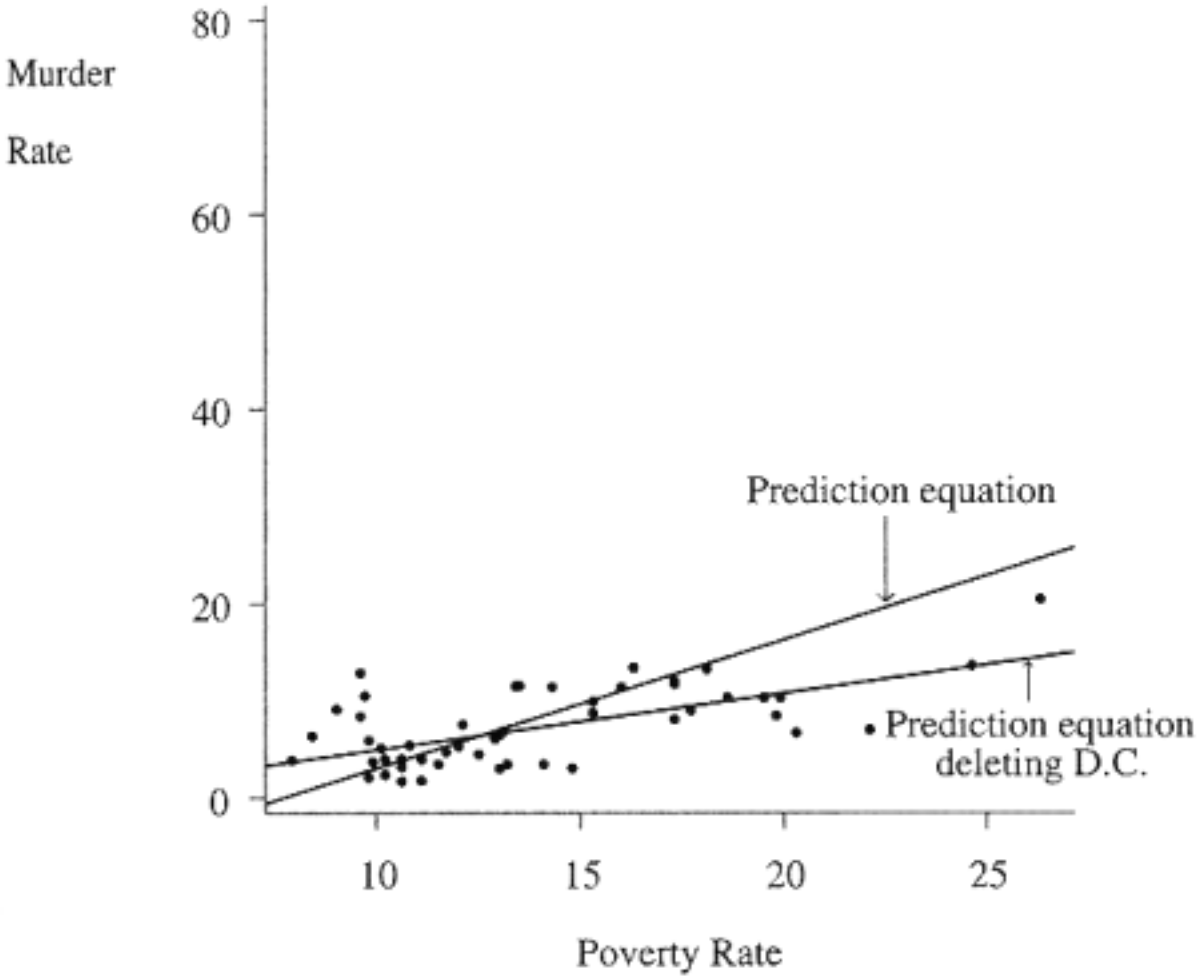

Figure 9.6 Prediction Equations Relating Murder Rate and Percentage in Poverty, with and without D.C. Observation

Graph the Distributions of the Dependent and Independent Variables

They should be roughly normally distributed.

If a given variable is not normally distributed, you may wish to consider transforming it (e.g., a log transformation).

If there is a serious violation of normality, you might consider dropping the variable from the analysis (e.g., if it is a independent variable, and not theoretically crucial).

You might also try another technique (e.g., nonlinear regression, logistic regression).

Some Tips for Transforming Data:

To correct **positive skews:**

- a **stronger** transformation: **$-1/X^2$**
- some **mild** transformations: **$-1/X$**

log X

$\sqrt{X}$

No shape change

Required: **X**

To correct negative skews:

- some **mild** transformations: $\mathbf{X^2}$

  $\mathbf{X^3}$

- a **stronger** transformation: **antilog X**

Plotting to test the assumptions of Multiple Regression Analysis

Examine the Scatterplots of each X by Y.

The assumption is that the bivariate relationship will be roughly linear.

If it is not linear, you might consider transforming one (or both) of the variables.

Examine the Plots of the Residuals by Each X.

The assumption is that the residuals are equally distributed at each value of X, and that the slope of the regression line $= 0$.

If not, you might consider a transformation, or a different form of the equation.

Examine the plot of the standardized residuals by the predicted value of Y.

The slope should be 0, the residuals spread out evenly at different levels of the predicted Y.

Examine the Normal P-P Plot.

This is the normal probability plot of the standardized residuals. It is used to check normality. If the variable is normally distributed, the plotted points form a straight diagonal line.

Use a histogram to examine the distribution of the residuals.

Similar to looking at the Normal P-P Plot.

The residuals should be roughly normally distributed. Keep an eye out in particular for severe outliers.

Checking for Constance Variance

It is assumed that variance is the same across all values of the independent variable.

Plot the standardized (or studentized residuals) against the predicted values. There should be equal spread, the slope should be 0).

Plot predicted Y by observed Y. Should be linear association, equal spread below and above the regression line.

Examine the partial plots of each X by Y.

Again, the assumption is that the partial relationship will be roughly linear.

If it is not linear, you might consider transforming one of the variables.

Independence of the Y's

It is assumed that the Ys are independent from one another. This may not be the case if data collection occurred over time, or if the dependent variable is somehow related to time.

You can check for potential problems of this nature by:

1. Examining the Durban Watson statistic.

   One of the assumptions of regression analysis is that the residuals for consecutive observations are uncorrelated. If this is true, the expected value of the Durbin_Watson statistic is 2. Values less than 2 indicate positive autocorrelation, a common problem in time_series data. Values greater than 2 indicate negative autocorrelation.

2. You can plot the residuals by the sequence of the observations.

Problems of Multicolinearity?

If an independent variable is strongly associated with another independent variable (colinearity, very high r^2), or if an independent variable is a strong linear function of the other independent variables in a regression model (multicolinearity, very high R^2) then problems may arise in the estimation of regression coefficients. Notably, multicolinearity causes inflated standard errors for estimates of regression coefficients, and can cause other problems (coefficients with the wrong sign, dramatic changes in the sign and size of a coefficient when another one is added to the equation.).

Detecting Colinearity/Multicolinearity

Examine the intercorrelation matrix. Very high values of r indicate a potential problem.

Examine the tolerance. For each independent variable, the tolerance is the proportion of variability of that variable that is not explained by its linear relationship with the other independent variables in the model ($1 - R^{2)}$. Tolerance can range from 0 to 1. A value close to 1 indicates that an independent variable has little of its variability explained by the other independent variables.

A value close to 0 indicates that a variable is almost a linear combination of the other independent variables. Tolerances of less than 0.1 may be a problem.

You can also look at the **Variance Inflation Factor** (VIF). This is the reciprocal of the tolerance. As the variance inflation factor increases, so does the variance of the regression coefficient, making it an unstable estimate. Large VIF values are an indicator of Multicollinearity.

Solutions to Multicolinearity

One solution is to omit a problem variable from the analysis.

Another, is if you have several variables that are conceptually related, and that are highly intercorrelated, then you might consider creating an index.

Examine the Casewise Statistics

Are there outliers?

If yes, how large are they (what is their standardized value?) If they are greater than 3.0 then they are unlikely to occur due to chance if the residuals are normally distributed.

Looking for Influential Points.

Influence Statistics

DfBeta(s): The difference in beta value is the change in the regression coefficient that results from the exclusion of a particular case. A value is computed for each term in the model, including the constant.

Standardized

DfBeta(s) Standardized differences in beta values. The change in the regression coefficient that results from the exclusion of a particular case. SPSS suggests that you may want to examine cases with absolute values greater than 2 divided by the square root of N, where N is the number of cases. A value is computed for each term in the model including the constant.

DfFit The difference in fit value is the change in the predicted value that results from the exclusion of a particular case.

Standardized

DfFit Standardized difference in fit value. The change in the predicted value that results from the exclusion of a particular case.

SPSS suggests that you may want to examine standardized values which in absolute value exceed 2 divided by the squared root of p/N where p is the number of independent variables in the equation and N is the number of cases.

Distances

Mahalanobis

A measure of how much a case's values on the independent variables differ from the average of all cases. A large Mahalanobis distance identifies a case as having extreme values on one or more of the independent variables.

Cook's

A measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients.

A large Cook's D indicates that excluding a case from computation of the regression statistics, changes the coefficients substantially.

Leverage Values

Measures the influence of a point on the fit of the regression. The centered leverage ranges from 0 (no influence on the fit) to (N-1)/N.