Diagnostics & Remedies in Simple Regression

1. Residual Analysis & the Healthy Regression

1. Residual Plot of the Healthy Regression

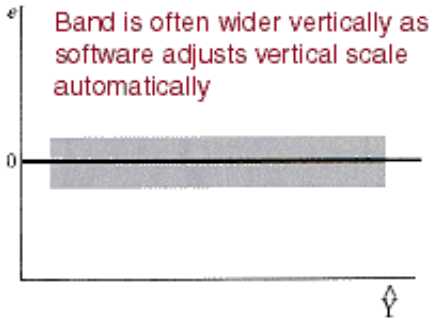
Residual analysis is a set of diagnostic methods for investigating the appropriateness of a regression model based on the residuals

$$e_i = Y_i - \hat{Y}_i$$

where $\hat{Y}_i$ is the predicted or fitted value (aka *predictor*, aka *estimate*)

$$\hat{Y}_i = b_0 + b_1 X_i$$

The basic idea is that, if the regression model is appropriate, the residuals e_i should "reflect the properties ascribed to the model error terms ε_i, such as independence, constant variance for different levels of X and normal distribution. The workhorse of residual analysis is the *residual plot*, or plot of the residuals e_i against the predicted values $\hat{Y}_i = b_0 + b_1 X_i$.



In a healthy regression the residuals should appear as arranged randomly in a horizontal band around the Y = 0 line as depicted to the right.

In reality, computer programs typically adapt the range of the data to occupy the entire frame of the plot, so that the residuals appear as a random cloud of points spread out over the entire vertical range of the plot.

2. Potential Problems with the Simple Regression Model

Problems with the regression are departures from (or *violations of*) the assumptions of the regression model. They are:

1. Regression function is not linear
2. Error terms do not have constant variance
3. Error terms are not independent
4. Model fits all but one or a few outlier observations
5. Error terms are not normally distributed
6. One or several important predictor variables have been omitted from model

These problems also affect multiple regression models. Most of the diagnostic tools used for simple linear regression are also used with multiple regression. We will examine these potential problems in conjunction with diagnostic tools which can be *informal* (such as graphs) or *formal* (such as tests), and remedies.

3. Using Informal (Graphic) Versus Formal Diagnostic Tests

Features of informal graphic diagnostics are

- they are visual and at least in part subjective
- they may be inconclusive
- they require judgment, and therefore
- they may require training and/or experience
- sometimes judgment may be helped by calibration (see normal probability plot, later)

Features of formal tests are

- they may give a straight yes/no answer about presence of a problem
- they may not require experience or training
- they may be used in automatic fashion
- they may not reveal the substantive situation as well as informal tools, so that
- researcher relying entirely on formal tools may miss an interesting aspect of the data (e.g., in Box-Cox estimation of optimal transformation of the data)

Researchers can use available informal and formal diagnostics to develop their own strategy to diagnose problems with their models.

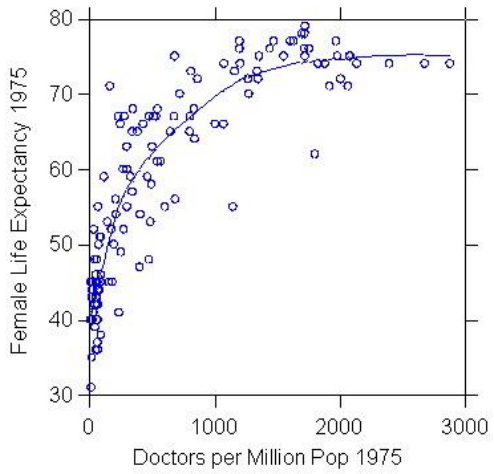
2. Regression Function is Not Linear

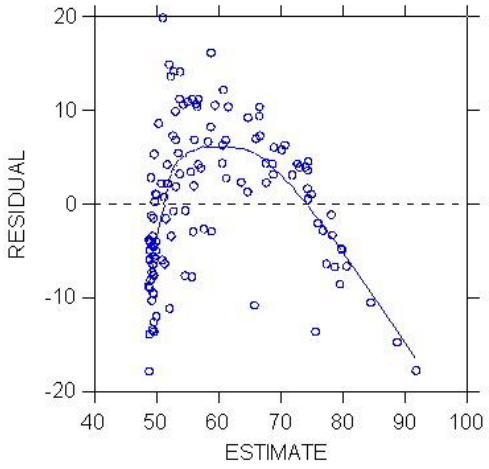
1. Scatterplot or Residual Plot with LOWESS Robust Nonparametric Regression Curve (Diagnostic, Informal)

Nonlinearity in the regression function may appear in the scatter plot of Y versus X, or in the residual plot. The residual plot often magnifies nonlinearity (compared to the scatter plot) and makes it more noticeable.

A nonlinear trend may be revealed by using a *nonparametric regression* technique such as LOWESS. The LOWESS algorithm estimates a curve that represents the main trend in the data, without assuming a specific mathematical relationship between Y and X. (This is why it is called *nonparametric*.)

The figures below show (1) a scatterplot of female life expectancy against doctors per million for 137 countries in 1975, and (2) a plot of the *residuals* of the regression of female life expectancy on doctors per million. Note how the nonlinearity is magnified by the residual plot, as compared to the plot of Y against X.





2. Linearity or Lack of Fit Test (Diagnostic, Formal, Limited Applicability)

There is a test of linearity of the regression function, called the *lack of fit* test. This test requires repeat observations (called *replications*) at one or more levels of X, so it cannot be performed with all data sets. In essence it is a test of whether the means of Y for the groups of replicates are significantly different from the fitted value $\hat{Y}_i$ on the regression line, using a kind of F test. The test is explained in the text pg. 119-127.

3. Polynomial Regression (Remedy)

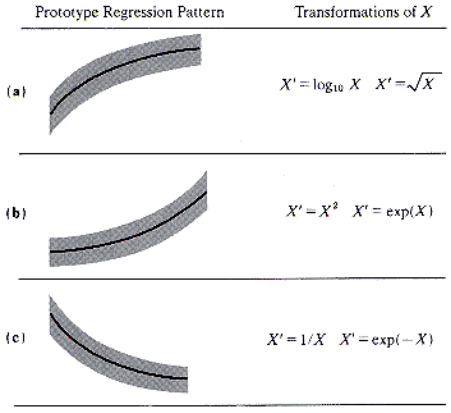
Some forms of non-linearity can be modeled by introducing higher powers of X in the regression equation. Polynomial regression is a form of multiple regression and is discussed in later sections.

4. Transforming Variables to Linearize the Relationship (Remedy)

1. Error Variance Appears Constant

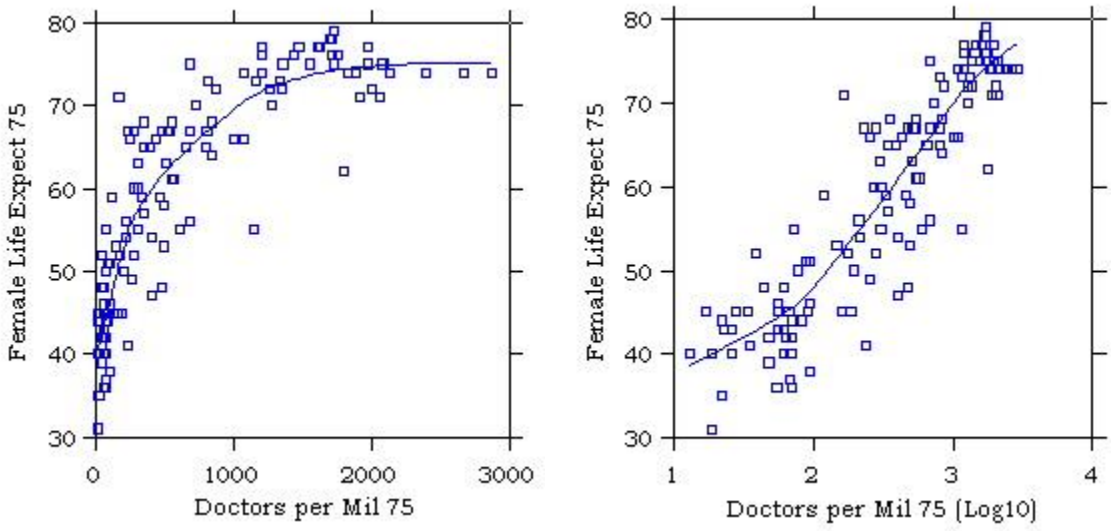
When *the error variance appears to be constant*, only X needs be transformed to linearize the relationship. Some typical situations are shown to in the figure to the right.

FIGURE 3.13 **Prototype Nonlinear Regression Patterns with Constant Error Variance and Simple Transformations of X.**



WHEN ERROR VARIANCE IS CONSTANT, ONE NEEDS ONLY TRANSFORM X, NOT Y

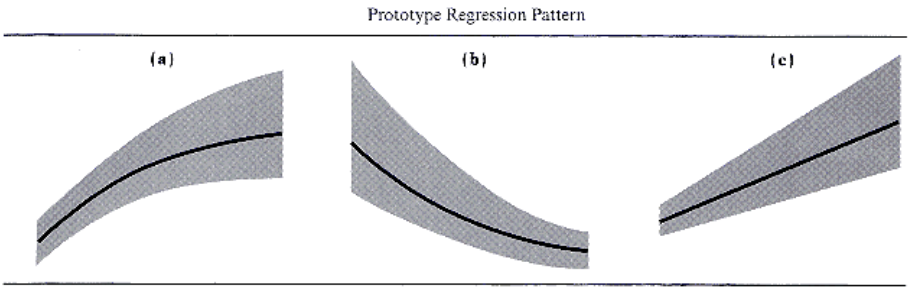
The following figures show an example of a non-linear regression function that can be straightened by transforming X.



2. Error Variance Appears Not Constant

When *the error variance does not appear constant* it may be necessary to transform Y or both X and Y. The next two exhibits show examples.

FIGURE 3.15 **Prototype Regression Patterns with Unequal Error Variances and Simple Transformations of Y.**



Note: A simultaneous transformation on X may also be helpful or necessary.

3. Transformation to Simultaneously Linearize the Relationship and Normalize the Distribution of Errors

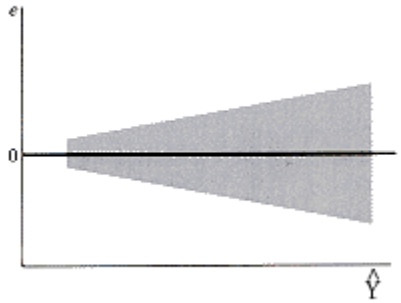
See Box-Cox transformation below.

3. Error Terms Do Not Have Constant Variance (*Heteroskedasticity*)

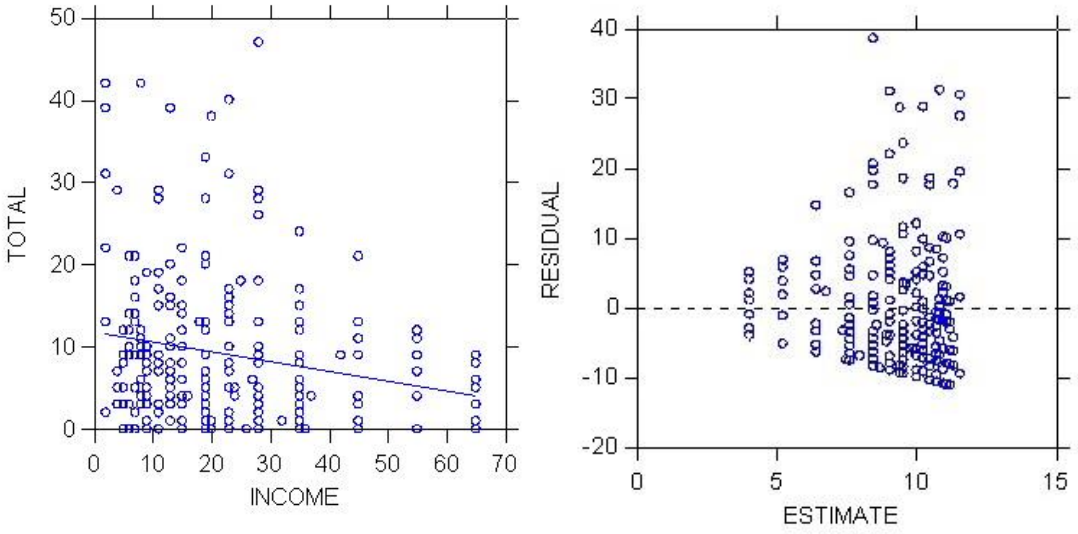
1. Funnel-Shape in in Residual Plot (Diagnostic, Informal)

Terminology:
homoskedasticity: σ^2 is constant over the entire range of X (as specified by OLS assumptions)
heteroskedasticity: σ^2 is not constant over the entire range of X (departure from OLS assumptions)

The regression model assumes homoskedasticity. Heteroskedasticity may be manifested by a funnel or megaphone pattern like the following prototype



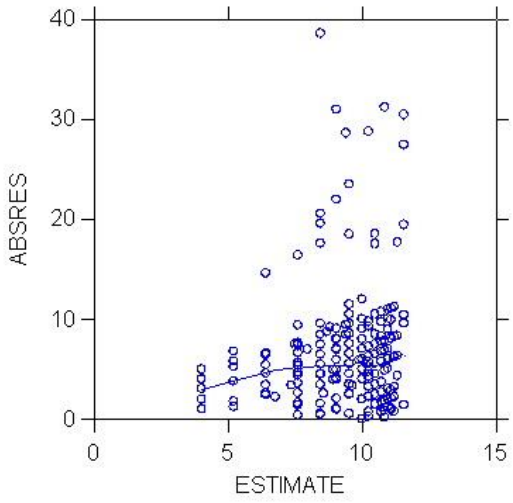
Sometimes the funnel pattern is reversed, with greater error variance corresponding to smaller values of X.



Q - Why is the variance in depression score lower for higher incomes and vice-versa? However, when the residual is plotted against $\hat{Y}_i$ the megaphone pattern fans out to the right.

2. Plot of Absolute Residual |e| or Squared Residual e^2 by $\hat{Y}$ (Diagnostic, Informal)

A pattern of heteroskedasticity can be detected informally by plotting the absolute residual or the squared residual against $\hat{Y}_i$; adding the linear fit or a LOWESS curve helps seeing the trend in the data. See figure to the right.



3. Tests for Homoskedasticity (Constancy of σ^2) (Diagnostic, Formal)

There are many tests of heteroskedasticity. Two formal tests of constancy of the error variance discussed in Chapter 3 are:

- the Brown-Forsythe Test (see pg. 116)
  - Robust to non-normal errors, i.e. does not depend on normality of error terms. Requires user to break data into two groups and test for constancy error variance across groups (not natural for continuous data).

- the Breusch-Pagan *aka* Cook-Weisberg Test (see pg. 118)
  - Tests whether the log error variance increases or decreases linearly with the predictor(s). Requires large samples & assumes normal errors.

4. Variable Transformation to Equalize the Variance of Y (Remedy)

See below on Tukey's ladder of powers and the Box-Cox transformation. Transformations of Y to make the distribution of residuals normal often have the effect of also equalizing the variance of Y over the entire range of X.

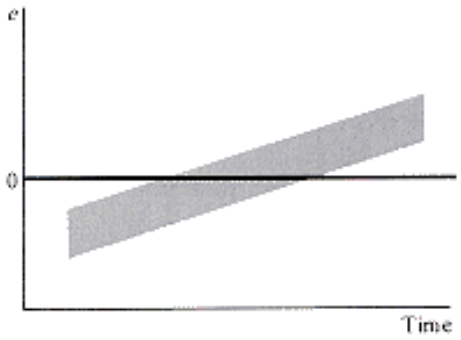
5. Weighted Least Squares (Remedy)

In rare cases where a variable transformation that takes care of unequal error variance cannot be found, one can use *weighted least squares*. In weighted least squares observations are weighted in inverse proportion to the variance of the corresponding error term, so that observations with high variance are downweighted relative to observations with low variance. Weighted least squares is discussed in Chapter 11.

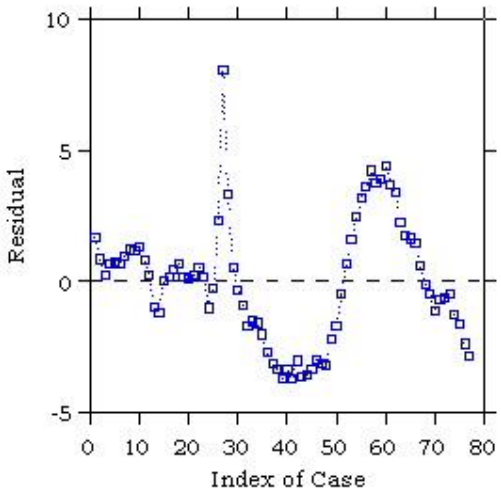
4. Error Terms Are Not Independent

1. Residual e by Time or Other Sequence (Diagnostic, Informal)

In time series data where observations correspond to successive time points, errors can be *autocorrelated* or *serially correlated*. Such a pattern can be seen in a plot of residuals against the time order of the observations, as in the figure to the right.



The following real example shows residuals for a regression of the divorce rate on female labor force participation (U.S. 1920-1996). Note the characteristic "machine gun" tracking pattern. The residual is plotted against the index of a case (corresponding to a year of observation).



Remedial techniques for lack of independence are discussed in Chapter 12.

2. Durbin-Watson Test of Independence (Diagnostic, Formal)

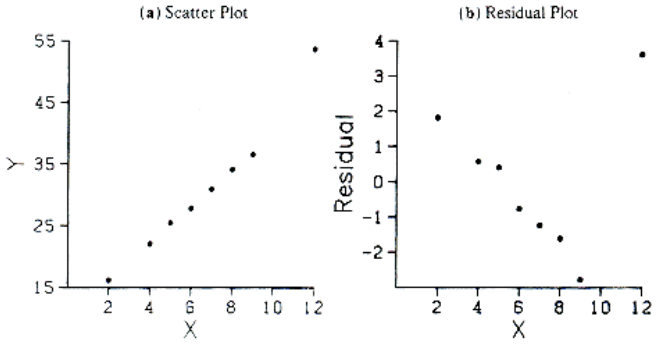
See Chapter 12.

5. Model Fits All But One or A Few Outlier Observations

1. Outliers in Scatterplot or Residual Plot (Diagnostic, Informal)

Outliers may often (but not always) be spotted in a scatterplot or residual plot.

FIGURE 3.7 Distorting Effect on Residuals Caused by an Outlier When Remaining Data Follow Linear Regression.



Outlying observations may affect the estimate of the regression parameters. Outlying observations may sometimes be identified in a box plot or stem-and-leaf display of the residuals.

To illustrate, use the applet:
http://www.shodor.org/interactivate/activities/Regression/

2. Tests for Outlier and Influential Cases (Diagnostic, Formal)

An extensive discussion of outliers and influential cases is provided in Chapter 10.

6. Error Terms Are Not Normally Distributed

Non-normality of errors corresponds to a range of issues with different degrees of severity. Kurtosis (fat tails) and skewness are more serious than more minor departures from the normal. These features of the distribution blend in with the problem of outlying observations.

1. Box Plot, Stem-and-Leaf, and Other Displays of the Distribution of e (Diagnostic, Informal)

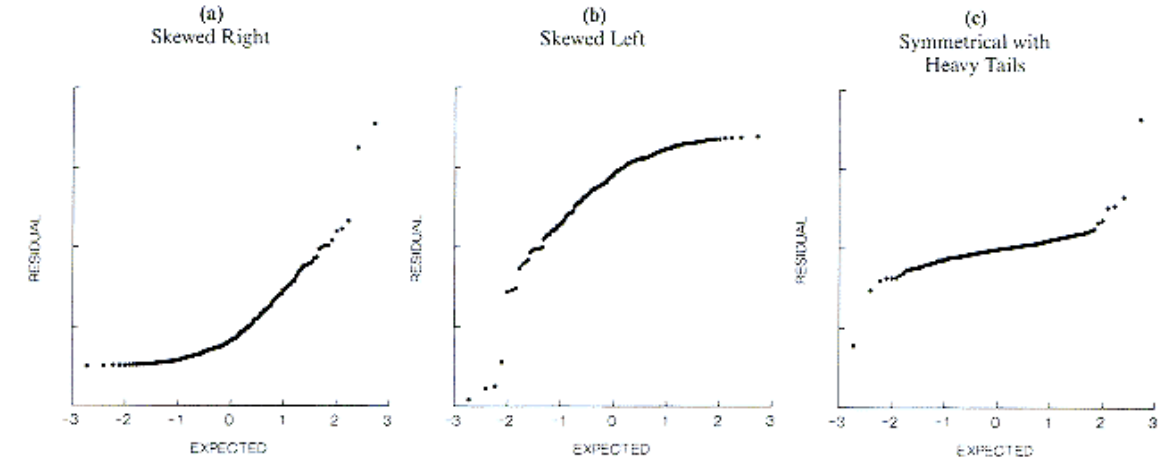
Lack of normality (as well as some types of outlying observations) may be diagnosed by looking at the distribution of the residuals using devices such as a histogram, stem and leaf display, box plot, kernel density estimator, etc. In the following examples see how the histogram of the distribution appears normal but fat tails/outlying observations are revealed in other displays.

2. Normal Probability (QQ) Plot of e (Diagnostic, Informal)

The normal probability plot is used to examine whether the residuals are normally distributed. It is a plot of the residuals against their expected values assuming normality. When the residuals are normally distributed, the plot is approximately a straight line.

The shape of normal probability plot reveals aspects of the distribution of e.

FIGURE 3.9 Normal Probability Plots when Error Term Distribution Is Not Normal.



3. Correlation Test for Normality (Diagnostic, Formal)

The correlation test for normality is based on the normal probability plot. It is the correlation between the residuals e_i and their expected values under normality. The higher the correlation, the straighter the normal probability plot, and the more likely that the residuals are normally distributed. The value of the correlation coefficient can be tested, given the sample size and the α level chosen, by comparing it with the critical value in Table B.6 in the textbook. A coefficient larger than the critical value supports the conclusion that the error terms are normally distributed.

4. Data Transformations Affecting the Distribution of a Variable (Remedy)

1. Standardizing a Variable - Does Not Affect Shape of Distribution

Common variable standardizations are the z-score and range standardization. These transformations do not affect the shape of the distribution of the variable (contrary to some popular beliefs).

2. Transforming a Variable to Look Normally Distributed

Transformations of Y can be used to remedy non-normality of the error term.

- Radical normalization with the rankit transformation
  - The rankit transformation transforms any distribution into a normal one. It is the same as *grading on the curve*. It is the same as finding the expected value of the residual in preparing a normal probability plot.
- Tukey's ladder of powers
  - Tukey (1977) has proposed a *ladder of powers* spanning a range of transformations to normalize the distribution of a variable in a data set.
  - The family of power transformations in Tukey's ladder is of the form $Y' = Y^{\lambda}$ where λ is the power parameter. Some of the principal steps of the ladder are

Lambda (λ)	Transformation	Name	SAS
2	$Y' = Y^2$	square	Y**2
1	$Y' = Y$	identity	Y
.5	$Y' = (Y)^{1/2}$	square root	sqrt(Y)
0	$Y' = \log(Y)$	logarithm	log(Y)
-.5	$Y' = 1/(Y)^{1/2}$	inverse square root	1/sqrt(Y)
-1	$Y' = 1/Y$	inverse	1/Y

  but λ can take any value in between.
- Automatic choice of ladder of powers transformation with the Box-Cox procedure
  - In the simplest version the Box-Cox procedure estimates the parameter λ by maximum likelihood so as to maximize the fit of the transformed data Y' to a normal distribution. There is only one variable involved.
- Box-Cox procedure to optimize the linear relationship between X and Y
  - The Box-Cox procedure can also be used to transform Y and X in such a way as to maximize their relationship. The procedure can estimate an exponent θ (theta) of X as well as the exponent λ (lambda) of Y. The complete model is thus $Y_i^{\lambda} = \beta_0 + \beta_1 X_i^{\theta} + \varepsilon_i$
  - The estimates λ, β_0 , β_1 and σ^2 can be found by maximizing the likelihood function (maximum likelihood has not been discussed)

7. One or Several Important Predictor Variables Have Been Omitted From Model

Plot of Residual e by Omitted Predictor Variable Z

Z is a potential predictor variable that is not included in the equation.