Chapter 3 Residuals (Residual Plots)

Residuals

The appropriateness of a simple linear regression model.

Recall that the four conditions that comprise the simple linear regression model are:

- The mean of the response , E(Yi), at each value of the predictor, xi, is a **Linear function** of the xi.
- The errors, ϵi, are **Independent**.
- The errors, ϵi, at each value of the predictor, xi, are **Normally distributed**.
- The errors, ϵi, at each value of the predictor, xi, have **Equal variances** (denoted $\sigma 2$).

An equivalent way to think of the first (linearity) condition is that the mean of the error, E(ϵi), at each value of the predictor, xi, is **zero**. An alternative way to describe all four assumptions is that the errors, ϵi, are independent normal random variables with mean zero and constant variance, $\sigma 2$.

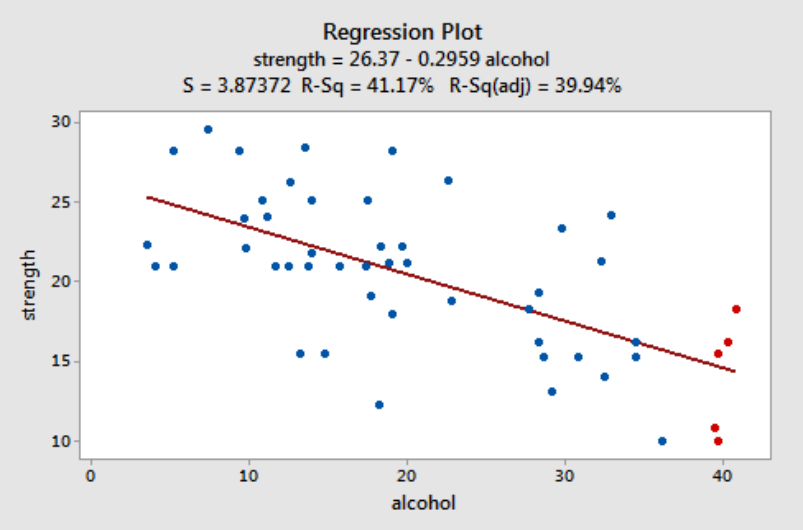
The four conditions of the model pretty much tell us what can go wrong with our model, namely:

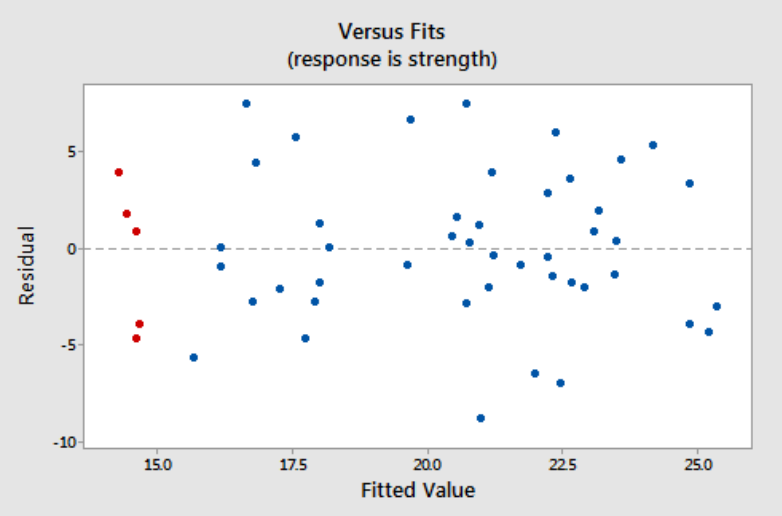
- The population regression function is **not linear**. That is, the response Yi is not a function of linear trend ($\beta 0+\beta 1xi$) plus some error ϵi .
- The error terms are **not independent**.
- The error terms are **not normally distributed**.
- The error terms do **not** have **equal variance**.

Chapter 3 Residual Analysis :

Using Residuals to access linearity of a model: Residuals versus Fitted Plot

When conducting a residual analysis, a "**residuals versus fits plot**" is the most frequently created plot. It is a scatter plot of residuals on the *y* axis and fitted values (estimated responses) on the *x* axis. The plot is used to detect non-linearity, unequal error variances, and outliers.





This plot is a classical example of a well-behaved residuals vs. fits plot. Here are the characteristics of a well-behaved residual vs. fits plot and what they suggest about the appropriateness of the simple linear regression model:

- The residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.
- The residuals roughly form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are equal.
- No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers.

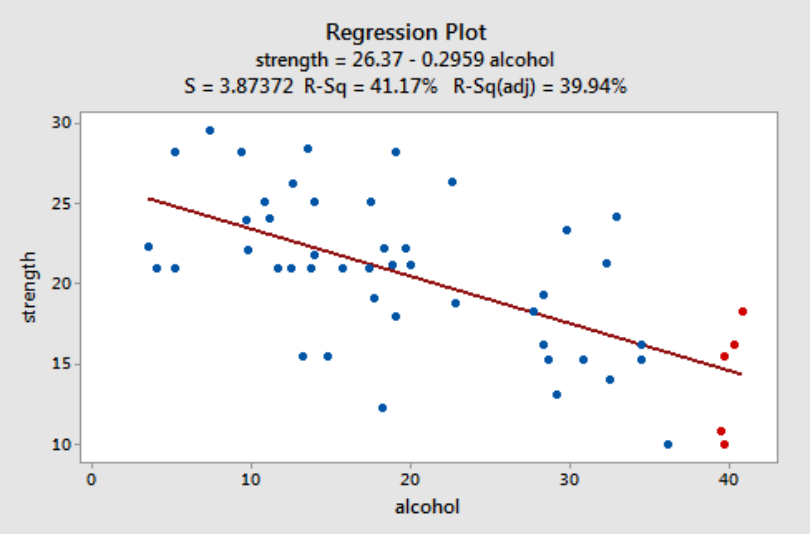
In general, you want your residual vs. fits plots to look something like the above plot

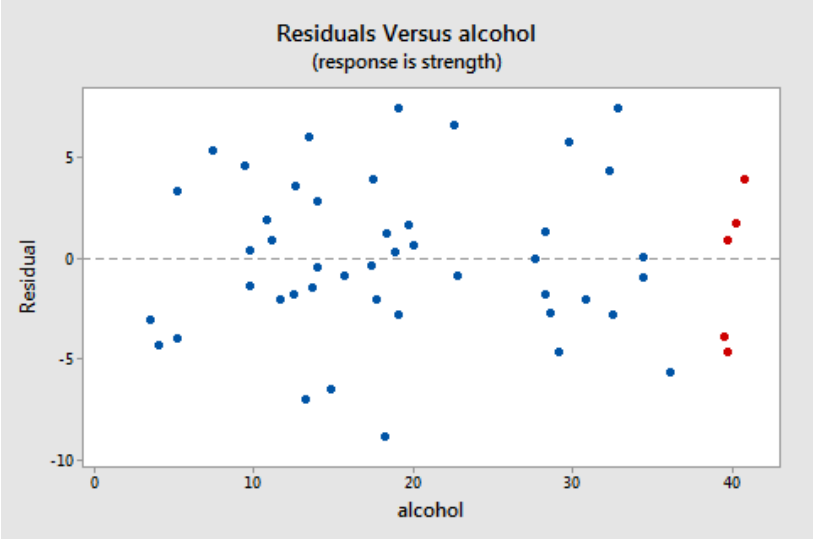
Chapter 3 - Residuals vs. Predictor Plot :

An alternative to the residuals vs. fits plot is a "**residuals vs. predictor plot**." It is a scatter plot of residuals on the *y* axis and the predictor (*x*) values on the *x* axis. For a simple linear regression model, if the predictor on the *x* axis is the same predictor that is used in the regression model, the residuals vs. predictor plot offers no new information to that which is already learned by the residuals vs. fits plot. On the other hand, if the predictor on the *x* axis is a new and different predictor, the residuals vs. predictor plot can help to determine whether the predictor should be added to the model (and hence a multiple regression model used instead).

The interpretation of a "residuals vs. predictor plot" is identical to that for a "residuals vs. fits plot." That is, a well-behaved plot will bounce randomly and form a roughly horizontal band around the residual = 0 line. And, no data points will stand out from the basic random pattern of the other residuals.

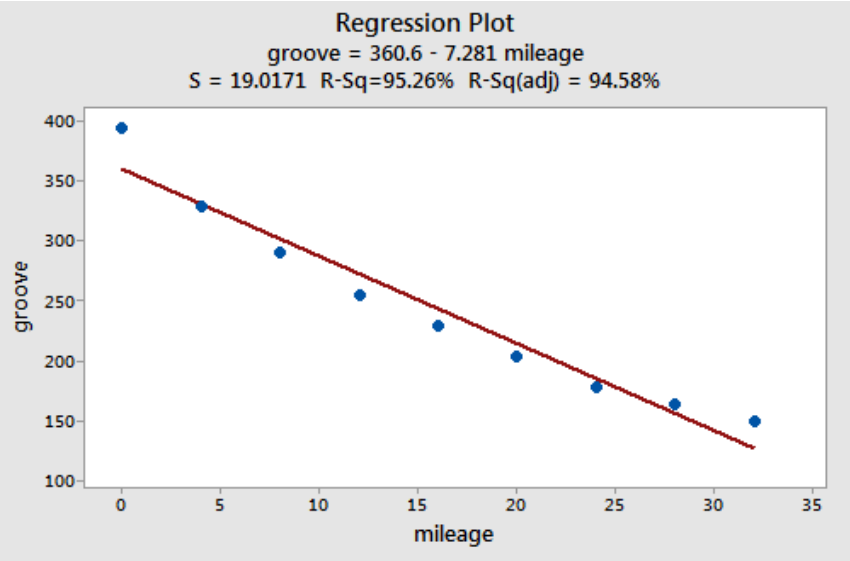
Here's the residuals vs. predictor plot for the simple linear regression model with arm strength as the response and level of alcohol consumption as the predictor:

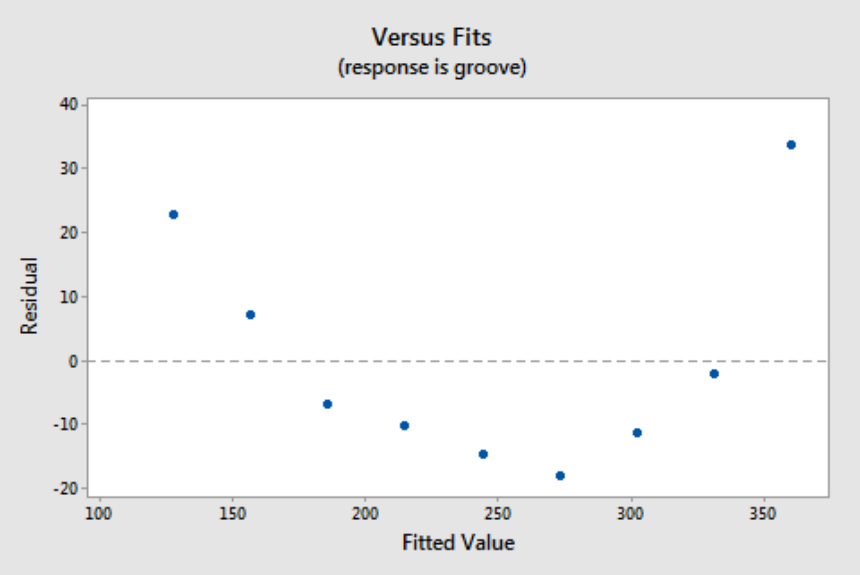




Identifying Specific Problems Using Residual Plots

Linear model not appropriate



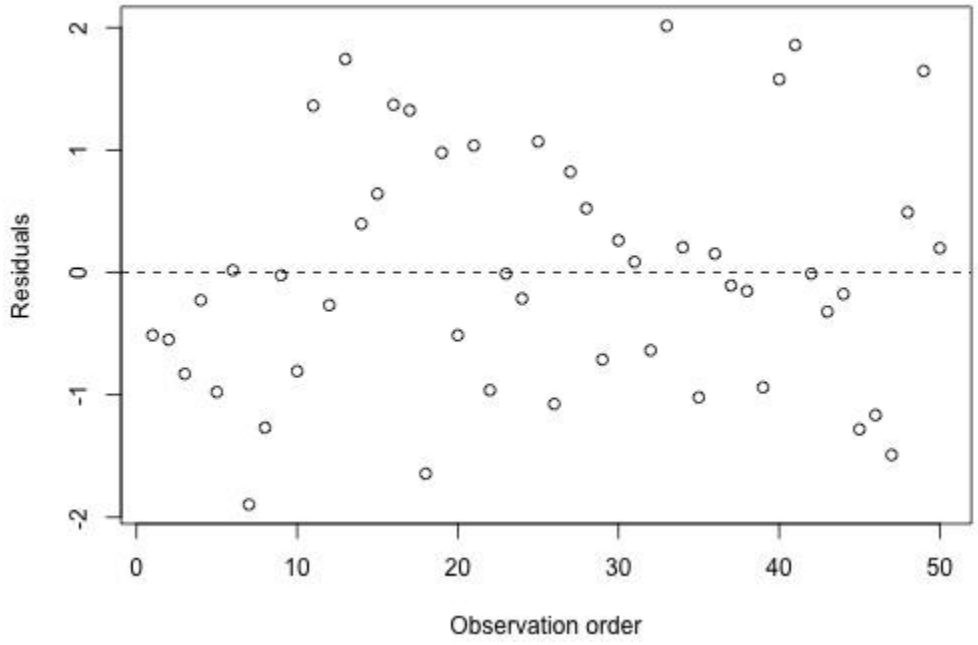


Residuals vs. Order Plot

Recall that the second condition — the "I" condition — of the linear regression model is that the error terms are independent. In this section, we learn how to use a "**residuals vs. order plot**" as a way of detecting a particular form of non-independence of the error terms, namely **serial correlation**. If the data are obtained **in a time (or space) sequence**, a residuals vs. order plot helps to see if there is any correlation between the error terms that are near each other in the sequence.

The plot is only appropriate if you know the order in which the data were collected! Highlight this, underline this, circle this, ..., er, on second thought, don't do that if you are reading it on a computer screen. Do whatever it takes to remember it though — it is a *very common* mistake made by people new to regression analysis.

So, what is the residuals vs. order plot all about? As its name suggests, it is a scatter plot with residuals on the *y* axis and the order in which the data were collected on the *x* axis. Here's an example of a well-behaved residuals vs. order plot:

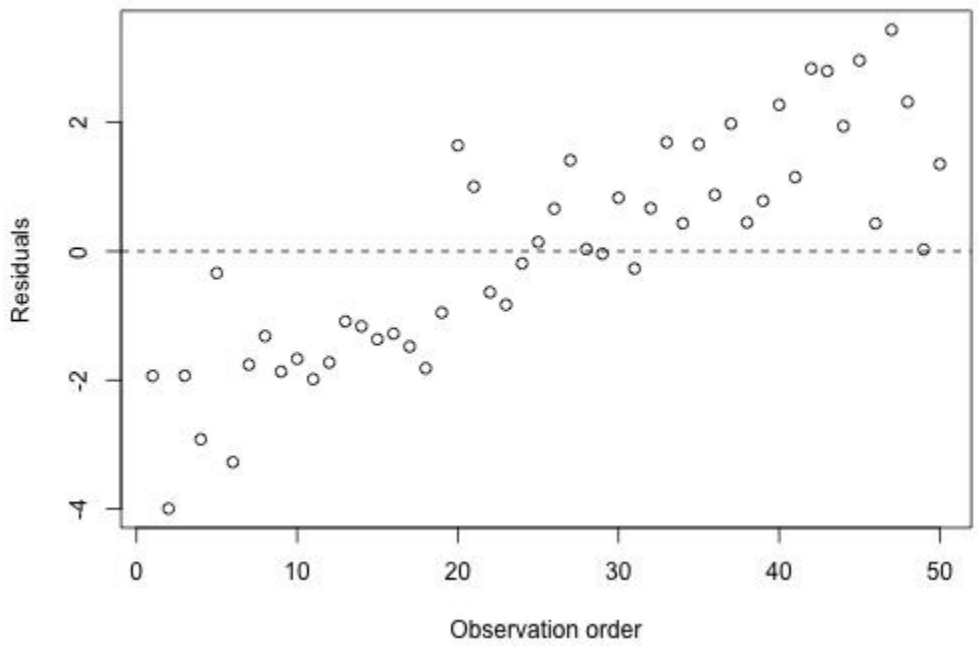


The residuals bounce randomly around the residual = 0 line as we would hope so. In general, residuals exhibiting normal random noise around the residual = 0 line suggest that there is no serial correlation.

Let's look at examples of the different kinds of residuals vs. order plots we can obtain and learn what each tells us.

A time trend

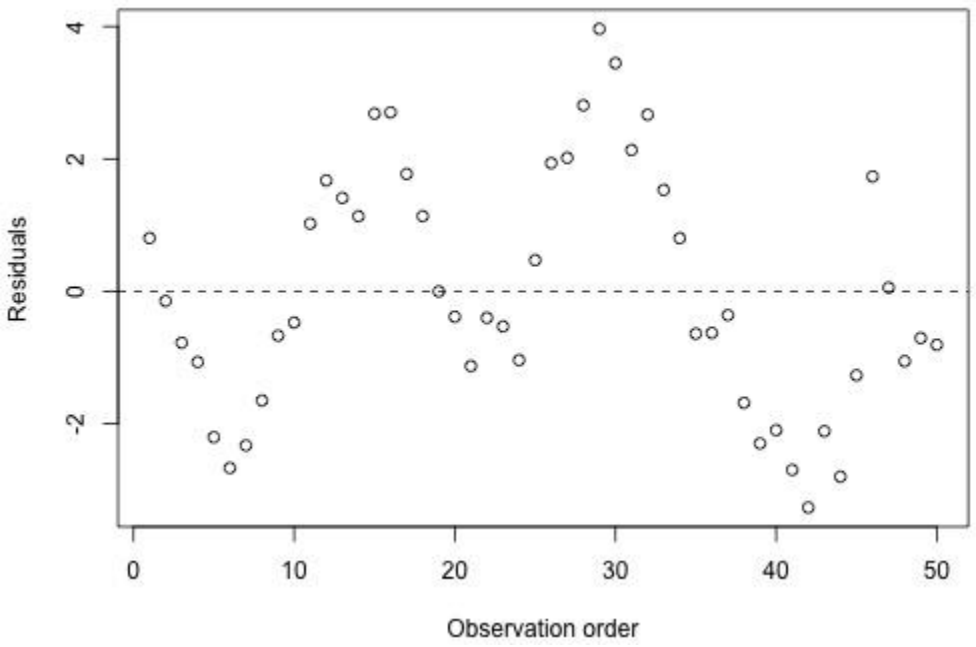
A residuals vs. order plot that exhibits (positive) trend as the following plot does:



suggests that some of the variation in the response is due to time. Therefore, it might be a good idea to add the predictor "time" to the model. That is, you interpret this plot just as you would interpret any other residual vs. predictor plot. It's just that here your predictor is "time."

Positive serial correlation

A residuals vs. order plot that looks like the following plot:



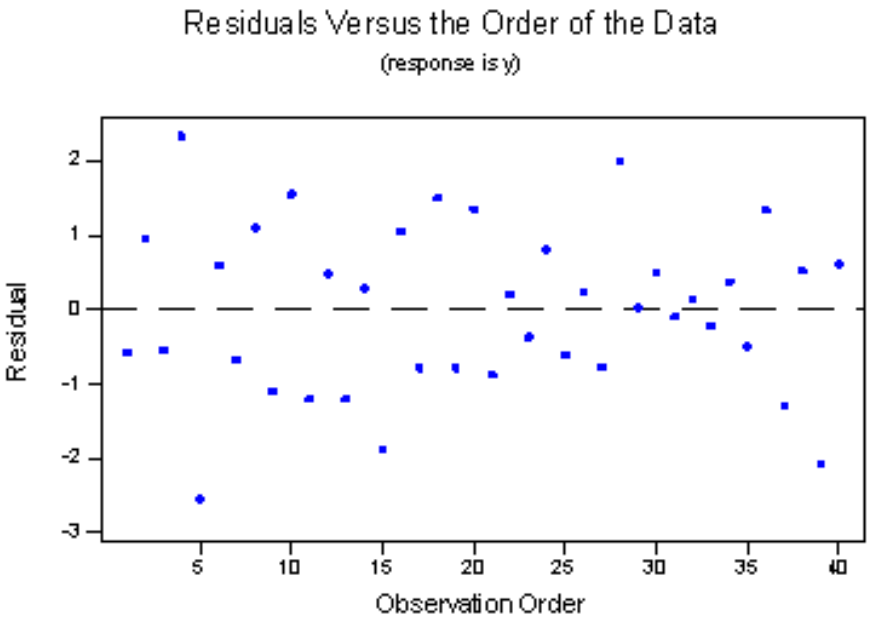
suggests that there is "**positive serial correlation**" among the error terms. That is, positive serial correlation exists when residuals tend to be followed, in time, by residuals of the same sign and about the same magnitude. The plot suggests that the assumption of independent error terms is violated.

Here is another less obvious example of a data set exhibiting positive serial correlation:

Can you see a cyclical trend -- up and then down, up and down, and up again? If not, **click on the "Draw trend!" icon**. Certainly, the positive serial correlation in the error terms is not as obvious here as in the previous example. These two examples taken together are a nice illustration of "the severity of the consequences is related to the severity of the violation." The violation in the previous example is much more severe than in this example. Therefore, we should expect that the consequences of using a regression model in the previous example would be much greater than using one in this example. In either case, you would be advised to move out of the realm of regression analysis and into that of "**time series modeling**."

Negative serial correlation

A residuals vs. order plot that looks like the following plot:



suggests that there is "**negative serial correlation**" among the error terms. Negative serial correlation exists when residuals of one sign tend to be followed, in time, by residuals of the opposite sign. What? Can't you see it? If you connect the dots in order from left to right, you should be able to see the pattern. If you can't see it, **click on the "Draw trend!" icon**:

Negative, positive, negative, positive, negative, positve, and so on. The plot suggests that the assumption of independent error terms is violated. If you obtain a residuals vs. order plot that looks like this, you would again be advised to move out of the realm of regression analysis and into that of "**time series modeling**."