Model Adequacy Checking

So far, we always made the following assumptions on the regression models we have fitted (without explicitly checking them every time):

- The relationship between each regressor and the response is approximately linear (we have looked at scatter plots to check that the relationship is not something other than a line, i.e., exponential or quadratic etc.)
- The residuals are
  - uncorrelated
  - Normally distributed
  - with mean zero (no need to check that for the least squares fit)
  - and constant variance σ^2.

We need the assumptions on the residuals to be approximately satisfied, so that we can use the methods we have derived so far for inference on the model parameters. If the residuals are not normally distributed, for instance, then the test statistic for testing whether a slope is zero *does not* have a t-distribution.

In practice, there is no guarantee that the assumptions are satisfied for any given data set. From the summary statistics, such as SS_R, SS_{Res}, t-and F-test statistics or R^2 it can be very hard (or impossible) to tell whether the assumptions are satisfied or violated. Thus we need other numerical or graphical methods to check the assumptions. The linearity is most often simply checked with scatter plots in which you look for obvious non-linear relationships between any of the predictor variables and the response. The remaining assumptions concentrate on the model residuals.

In the following, let ϵ denote the theoretical residuals (on which we can make distribution assumptions) and e the values of the residuals computed from the data.

Residual Analysis

The residuals can be computed after the model is fitted via the equations

$$e_i = y_i - \hat{y}_i, \qquad i = 1, \ldots, n$$

They describe the variation in the response variable that the linear model is *not* able to explain. We can think of the residuals as observations taken on a random variable

$$\epsilon \sim N(0, \sigma^2)$$

The residuals will always have sample mean zero by the way they are computed (recall that $\sum y_i = \sum \hat{y}_i$). Their variance is estimated by

$$MS_{Res} = \frac{SS_{Res}}{n-k-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}$$

where $k+1$ is the number of β-parameters in the model. Technically, the residuals are not independent (as they sum to zero, if you know $n-1$ of them, you know the n^{th} one). In fact, they have only $n-k-1$ degrees of freedom associated with them. If $k+1$ is (much) smaller than n, this non-independence of the residuals has no big negative effect on the model.

Scaling Residuals

Scaling residuals can be helpful in identifying OUTLIER observations. Outliers are observations that are not conform with the pattern exhibited by the rest of the data.

STANDARDIZED RESIDUALS: Since the mean of the residuals is zero, and their estimated variance is MS_{Res}, the quantities

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}, \qquad i = 1, \ldots, n$$

have approximately a standard Normal distribution. Large standardized residuals (e.g., $|d| > 3$) can be used to identify outlier observations.

STUDENTIZED RESIDUALS: Actually, the estimate MS_{Res} is only an approximation for the variance of each residual. In reality, the residuals *do not* all have the same variance. Their variance depends on the predictor values. Generally, residuals in the center of the predictor variables tend to be larger than residuals at more remote locations. This is the opposite of what happens with predictions of new observations. The estimated least squares regression line tends to accommodate the observed extreme points more than the points in the center.

We can obtain the exact standard deviation of the i^{th} residual by using the relationship

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ was the "hat"-matrix.

Fact: The "hat" matrix $\mathbf{H}$ is SYMMETRIC ($\mathbf{H}' = \mathbf{H}$) and IDEMPOTENT ($\mathbf{HH} = \mathbf{H}$).

As the "hat" matrix maps the actual observations $\mathbf{y}$ onto their predicted values $\hat{\mathbf{y}}$, it can also be used to map the theoretical residuals onto the observed residuals:

Thus, the actual covariance matrix of the residuals is:

The covariance matrix of $\mathbf{e}$ is generally not diagonal, and the diagonal elements are not all the same. That means that the residuals are correlated and that they do not all have the same variance. Using MS_{Res} to estimate the variance of each residual actually overestimated those variances.

Def: The STUDENTIZED RESIDUALS are defined as

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}, \qquad i = 1, \ldots, n$$

They all have constant variance $Var(r_i) = 1$ regardless of the location of $\mathbf{x_i}$, provided the model that has produced them is correct.

PRESS RESIDUALS: PRESS stands for PRediction Error Sums of Squares. The idea behind this approach is to check whether the i^{th} observation is an outlier, by leaving the i^{th} observation out when fitting the regression line and comparing the fit. The PRESS residuals are $y_i - \hat{y}_{(i)}$ where $\hat{y}_{(i)}$ is the predicted value of observation i based on the regression that did not use observation i. The big idea here is that if an observation is very influential, leaving out this observation could drastically change the appearance of the regression line and would thus cause a much larger PRESS residual than ordinary residual. This calculation is repeated for each observation in the model. Sometimes, these residuals are also referred to as DELETED RESIDUALS.

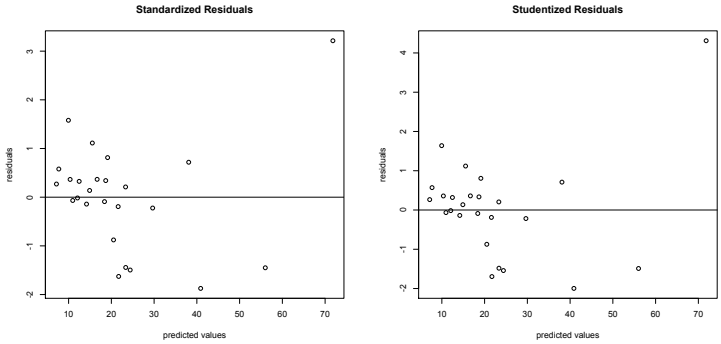
It is possible to show that the PRESS residuals $e_{(i)}$ are related to the ordinary residuals e_i via the equation

$$e_{(i)} = \frac{e_i}{1-h_{ii}}, \qquad i = 1, \ldots, n$$

The standardized PRESS residuals are identical to the Studentized residuals.

To quickly gain an overview of the magnitude of the standardized or Studentized residuals, the residuals are often plotted against the fitted values of the response $\hat{y}_i$ or against the individual predictors in order to identify outliers or influential points.

Example: After a linear model `fit` is fit in R, the standardized and Studentized residuals can be obtained with the commands `rstandard(fit)` and `rstudent(fit)`, respectively.



Residual Plots

Graphical analysis is much more effective in trying to detect patterns in the residuals than looking at the raw numbers. There are different types of plots that can be employed to check the different model assumptions.

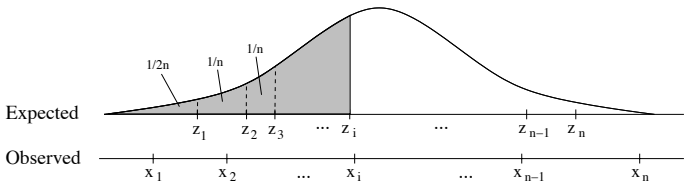
CHECKING NORMALITY: Recall, that one model assumptions is that the residuals have a Normal distribution. In reality, residuals will never be perfectly normally distributed, but we have to judge how grave the non-normality is for any specific data set. If the departure from normality is drastic, then the test statistics that we use to test anything (for instance whether slopes are equal to zero) *do not* have a t-distribution. Thus, the p-values that software computes for these tests become entirely meaningless.

A good way to check whether data has any specific distribution is with so-called PROBABILITY PLOTS (pp-plots) or with QUANTILE PLOTS (qq-plots). R produces qq-plots very easily and pp-plots only with intensive coaxing.

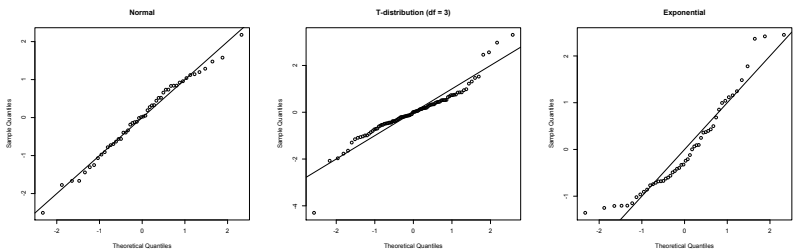
The basic idea behind a qq-plot:

1. Let $F(x)$ denote the CDF of the hypothesized distribution (e.g., Normal). Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote the actually observed data - sorted by size. $x_{(1)}$ is the smallest observation and $x_{(n)}$ is the largest.

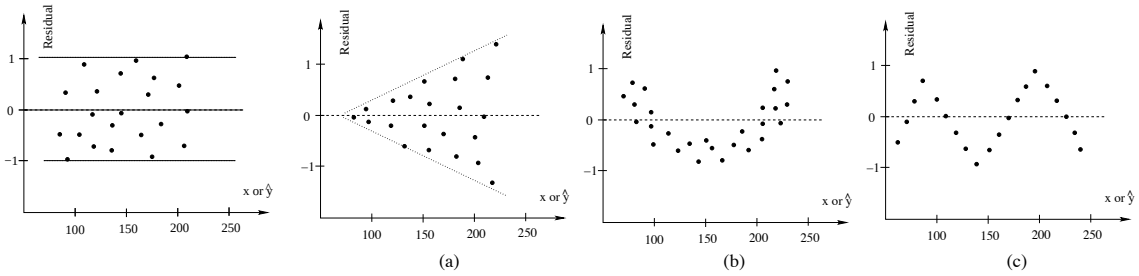
2. Compare the sample quantiles $x_{(i)}$ to the theoretical quantiles $z_i = F^{-1}(\frac{1-0.5}{n})$ of data that is ideally spaced out across the hypothesized distribution.



If your data comes from the hypothesized distribution, then the points in the qq-plot should lie on (or close to) the line $y = x$. If the points do not lie on the $y = x$ line, then the distribution of the data in not Normal. With a little practice, you can tell from a qq-plot, *how* the data deviates from the normality assumption (tails too thick, tails too thin, skewed etc.).



We have seen previously, that it can be helpful to plot the computed (and possibly standardized or Studentized) residuals against the regressors or predicted values to identify outliers. But these types of residual plots have another purpose. Recall, that one of our model assumptions was that the variance of the standardized or Studentized residuals should be same (regardless of the values of the predictors or response). If this assumption is not satisfied it can usually be seen in the residual plot.



In the panel on the left, the residuals are distributed "like confetti in a box". If this is the case, then the assumption on constant variance is approximately satisfied. The right three panels show us different problems with the data. In (a), the residual variance is clearly not constant, it increases, as the value that the residual is plotted against increases. In (b), the residuals seem to be a function (here quadratic) of the variable they are plotted against. In (c), the residuals are dependent on each other (and thus not independent). Problems like those exhibited in (a) or (b) can often be corrected with a variable transformation on either the predictors, or the response, or both. If the residuals are dependent on each other, we say that there is a significant AUTOCORRELATION between them. This is a serious violation of the regression assumptions that frequently occurs in situations where one of the predictors has a sequential structure (such as time) and should be addressed. Chapter 15 of your text describes methods with which this can be done.

PRESS Statistic

Recall, that the PRESS residuals (also known as deleted residuals) are the ones obtained from fitting a regression model with a single observation deleted. Large PRESS residuals are helpful in identifying observations for which the model does not fit the data well. The sum of the squared PRESS residuals can be used as a measure of the quality of a regression model (small is good).

$$\text{PRESS} = \sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

This statistic is useful in judging how well the model will predict new data. To normalize the PRESS statistic and make it useful in comparing models, an R^2 for prediction based on PRESS is computed as

$$R^2_{\text{prediction}} = 1 - \frac{\text{PRESS}}{SS_T}$$

This R^2 can be interpreted as the percentage of variability in predicting new observations that the model is able to explain. In particular, this criterion is useful when comparing models for the same data set to each other.

Example: For the Delivery Time data, the PRESS statistic for the full model is 459.0393, which makes

$$R^2_{\text{prediction}}\ (FM) = 1 - \frac{459.0393}{5784.5} = 0.9206$$

Compare this value to the PRESS statistic and $R^2_{\text{prediction}}$ value for the model that contains only the predictor CASES:

$$\text{PRESS} = 733.55, \qquad R^2_{\text{prediction}} = 1 - \frac{733.55}{5784.5} = 0.8732.$$

Since the value of $R^2_{\text{prediction}}$ for the one-predictor model is quite a bit smaller than for the two-predictor model, this is an indication that the variable DISTANCE should be included in the model to reliably model the delivery time.

Outliers in Regression

Outliers are values that are not "typical" responses for a given data set. They can be caused by measurement or recording errors in a single observation, or they could be cases for which the same explanation (model) as for the rest of the data does not hold. There are two possible treatments for outliers: removing them from the data and fitting a model without them, or trying to change the model to accommodate the unusual observation.

If it can be determined that the outlier is the result of an error (measurement, recording etc.), then removal may be the best option. But in some cases, outliers actually carry very useful information. So outlier points should only be removed from a data set with caution and usually only after you discuss their removal with the experimenter. Depending on where the outlier lies in $\mathbf{x}$-space, it can have a moderate or strong influence on the fitted regression model. The influence is more pronounced, the further away the outlier is from the center of the $\mathbf{x}$-space.

We have already briefly discussed several options for detecting outliers in a given data set:

- Looking for large (magnitude > 3) standardized residuals
- Looking at the PRESS residuals and comparing them to the regular residuals
- Plotting residual plots and looking for observations far away from the "bulk"

In any real problem it is advisable to use all three methods to identify INFLUENTIAL POINTS (points whose removal from the model would cause a large change in the estimated model parameters). These points may or may not be outliers. If you can

be sure that an error has been made, you should exclude the point(s) from the data set. If you cannot communicate with the experimenters, that's usually not possible. Then it is a good idea to check how large of a change the exclusion of influential points would have on the model as measured by

- The estimated regression parameters
- R^2
- MS_{Res}
- The standard errors of the slope(s) in the model

In theory, you would want high R^2, low MS_{Res} and low standard errors for the slope(s). If the model in which the influential points are removed is much improved over the original model, exclusion of the points may be warranted.

Exception: If there are two mechanism at work in your data and your "outlier" is a product of a different mechanism than the rest of the data, exclusion of that data point would ignore one of the two mechanisms. Whenever it is possible to communicate with the original experimenters, it is advisable to bring potential influential values to their attention before removing them and asking for explanations that could then be used to improve the model.

Lack of Fit in Regression

So far, we have seen ways in which we can convince ourselves that the model assumptions on the residuals are satisfied. Another assumption in a linear model is that the relationship between the predictor and the response is indeed linear. Suppose that the residuals are independent, Normally distributed and have constant variance. Then there exists a formal test that answers the linearity question, provided that there are replicated observations on the response for at least one level of the predictor x.

NOTATION: Suppose that we have n_i observations on the response at the i^{th} level of the predictor $x_i, i = 1, \ldots, m$. Let y_{ij} denote the j^{th} response measurement at predictor level x_i.

Big Idea: Partition the residual sum of squares into two components: One caused by PURE ERROR - e.g., measurement uncertainty; and one caused by the lack-of-fit.

$$SS_{Res} = SS_{PE} + SS_{LOF}$$

Partition the residual sum of squares accordingly

$$\sum_{i=1}^{m}\sum_{i=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{i=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

If the constant variance assumption of the model is correct, then SS_{PE} is a model-independent measure of the pure error with degree of freedom $n - m$. We need one degree of freedom at each predictor level to estimate the error variance. The sum of squares for the lack of fit is a weighted sum of the deviations of the group response means from the estimated regression values. If SS_{LOF} is small, the fit of the model is good. SS_{LOF} has $m-2$ degrees of freedom, because we need to estimate the slope and intercept to estimate the $\hat{y}_i$.

The test statistic for the official lack-of-fit test is

$$F = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}} \sim F_{m-2,n-m}$$

If the null hypothesis is true (H_0: true regression function is linear) then this test statistic has an F-distribution with $m - 2$ and $n - m$ degrees of freedom. If the value of the test statistic is large (larger than the appropriate quantile), then we reject the null hypothesis and conclude that the relationship between the predictor and the response is nonlinear.

Note: If the data does not have repeated observations at the exact same level of the predictor x, then "close neighbors" can be used to estimate the pure error variance of the model. Sliding window approaches are often used to accomplish this task.