Spring 2015 Statistics 151 (Linear Models) : Lecture Fifteen

Aditya Guntuboyina

17 March 2015

1 Regression Diagnostics

The material for this lecture is taken almost verbatim from Chapter 4 of Julian Faraway's book on linear models with R. The estimates for β and their confidence intervals etc in the linear model depend on the assumptions underlying the linear model. In particular, note that we have assumed:

1. The errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2 and are normally distributed.
2. We have assumed that the expected value of the response vector Y equals $X\beta$.
3. We have assumed that all the subjects obey the same linear model. In practice, it may happen that a few subjects do not obey the model. These few observations might change the choice and fit of the model.

Fortunately, we can do regression diagnostics to check if the data show any evidence of deviating from the assumptions. Regression diagnostics should always be performed after regression analysis.

1.1 Checking Assumptions on $e_1, \ldots, e_n$

The errors $e_1, \ldots, e_n$ are of course unobservable. How does one then check the assumptions of independence, constant variance and normality of the errors? The idea is to use the residuals $\hat{e_1}, \ldots, \hat{e_n}$ which act as proxies for the errors. It is important to note that the residuals are not exactly interchangeable with the errors however. For example, $var(\hat{e_i}) = \sigma^2(1 - h_{ii})$ where h_{ii} is the ith leverage and $cov(\hat{e}_i, \hat{e}_j) = -\sigma^2 h_{ij}$ where h_{ij} is the (i, j)th entry of the hat matrix.

Because each h_{ii} lies between 0 and 1 and their average is $(1 + p)/n$ which is usually small, in most of the cases, it turns out that each h_{ii} is small. Hence each $\hat{e}_i$ has roughly variance equal to σ^2. Similarly, because $\sum_{j=1}^{n} h_{ij}^2 = h_{ii}$ for each i, it also turns out that h_{ij} is typically close to zero for most i and j. Thus the residuals $\hat{e}_1, \ldots, \hat{e}_n$ have variance roughly equal to σ^2 and correlation roughly equal to zero. This is true under the assumption that $e_1, \ldots, e_n$ are independent with variance σ^2. The residuals can therefore be used to test assumptions on $e_1, \ldots, e_n$. Alternately, one might use standardized residuals.

1.1.1 Constant Variance Assumption

Plot residuals (y-axis) against the fitted values (x-axis). If all is well, you should see constant variance in the vertical direction and the scatter should be symmetric vertically about zero. Things to look for are heteroscedasticity (nonconstant variance) and nonlinearity (which indicates that some change in the model is necessary).

Also plot the residuals (y-axis) against each explanatory variable values (for explanatory variables that are both in and out of the model; we will be looking at variable selection methods later). Look for the same things as the residuals against fitted values plot; except that in the case of plots against explanatory variables that are not in the model, look for any relationship that might indicate that this explanatory variable should be included.

If indeed there is some evidence of nonconstant variance, two common ways of dealing with it are (a) using weighted least squares and (b) using variable transformations. We will look at weighted least squares later. The most common variable transformations are taking powers (most common power is square root) and logarithms. Some heuristic for these transformations is given below.

Suppose y is a random variable with mean μ. For a function $h(y)$, using a Taylor expansion of order one of y around μ, we get $h(y) \approx h(\mu) + h'(\mu)(y - \mu)$. From here, we obtain that $var(h(y)) \approx (h'(\mu))^2 var(y)$. Thus if $var(y) \propto \mu^2$, then use $h(y) = \log y$. If $var(y) \propto \mu$, use $h(y) = \sqrt{y}$.

Note however that a square root or logarithm can only be taken for nonnegative data.

1.1.2 Normality

Normality of the errors is checked by checking normality of the residuals. This is done via a qq plot. The qq-plot of the residuals plots the sorted residuals against $\Phi^{-1}(i/(n+1))$ for $i = 1, \ldots, n$.

When the errors are not normal, least squares estimators may not be optimal (although they are still best linear unbiased estimators). Other robust estimators may bemore effective. More importantly, tests and confidence intervals are not exact. However, only long-tailed distributions cause large inaccuracies. Mild nonnormality can safely be ignored and the larger the sample size the less troublesome the nonnormality.

The resolution of nonnormality depends on the type of problem found. For short-tailed distributions, the consequences of nonnormality are not serious and can reasonably be ignored. For skewed errors, a transformation of the response might solve the problem. For long-tailed errors, we might just accept the nonnormality and base the inference on the assumption of another distribution or use resamplign based methods such as permutation tests or bootstrap. Alternatively, one may use robust methods which give less weight to outlying observations.

Also you may find that other diagnostics suggest changes to the model. In this changed model, the problem of nonnormal errors might not occur. The Shapiro-Wilk test is a formal test for normality where a small p-value indicates non-normality. We will not go into the details of this test.

1.1.3 Correlated Errors

This is only problematic if there is a natural time (or spatial) structure to the way in which data on the subjects are collected. The simplest way of assessing correlation is to plot the sample autocorrelation function of the residuals (or standardized residuals).

The sample autocorrelation of the residuals $\hat{e}_1, \ldots, \hat{e}_n$ at lag k is defined by

$$\rho_k := \frac{\sum_{t=1}^{n-k} \hat{e}_t \hat{e}_{t+k}}{\sum_{t=1}^{n} \hat{e}_t^2}.$$

If there is no correlation structure in the residuals, we expect $\rho_1, \rho_2, \ldots$ to behave like independent normal random variables with mean zero and standard deviation $n^{-1/2}$. In R, one can plot the sample autocorrelations by the function $acf()$. This plot also gives two horizontal blue bars at the levels $\pm 1.96 n^{-1/2}$. If about 95% of the sample autocorrelations $\rho_1, \rho_2, \ldots$ (note that ρ_0 is always equal to 1) lie between the horizontal blue bars, then one need not worry about the errors being correlated.

There is also a formal test for checking correlation between errors. This is the Durbin-Watson test. We won't go into the details of this test.