6 Diagnostics and transformations for multiple linear regression

Assumptions:

1. main tool: standardized residuals and fitted values;
2. leverage points
3. outliers
4. added variable plots
5. collinearity-variance inflation factors

Recall multiple linear regression model and derive that Hat matrix

$$H = X(X'X)^{-1}X'$$

and $\hat{Y} = HY$. Let h_{ij} be the ijth element in the Hat matrix. $\hat{y}_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i}^{n} h_{ij}y_j$

Rules for identifying leverage points:

$$h_{ii} > 2 * average(h_{ii}) = 2\frac{p+1}{n}$$

Properties of residuals in multiple regression

$$E(\hat{e}|X) = 0$$

$$V(\hat{e}|X) = \sigma^2(I - H)$$

Note $HH' = H^2 = H$. Standardized residuals

$$r_i = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}}$$

where $s = \sqrt{\frac{1}{n-(p+1)}\sum_{j=1}^{n}\hat{e}_j^2}$ is the usual estimate of σ.

Common practice of labelling points as outliers in small to moderate size data sets, if the standardized residual for the point falls outside the interval from -2 and 2; for very large data sets, we shall change this rule to -4 to 4.

Using residuals and standardized residuals for model checking

When a valid model has been fit, a plot of the standardized residuals against any predictor

or any linear combination of the predictors will have the following features:

1. A random scatter of points around the horizontal axis.
2. Constant variability as we look along the horizontal axis.

An implication of these features is that any pattern in a plot of standardized residuals is indicative that an invalid model has been fit to the data.

In multiple linear regression, plots of residuals or standardized residuals provide direct information on the way on which the model is misspecified when the following two conditions hold:

$$E(Y|X = x) = g(\beta_0 + \beta_1 * x_1 + ... + \beta_p x_p) \tag{1}$$

$$E(X_i|X_j) \approx a_0 + a_1 X_j \tag{2}$$

This finding is based on the work of Li and Duan (1989). The linearity condition, (2) is another way to say that the predictors follow a elliptically symmetric distribution. Note that the X's follow a multivariate normal distribution then this is stronger than the condition for elliptically symmetric distribution. If either of the above condition does not hold, then a pattern in a residual plot indicates that an incorrect model has been fit, but the pattern itself does not provide direct information on how the model is misspecified. For example, we shall display nonconstant variance when the errors in fact have constant variance but the conditional mean is modelled incorrectly. Cook and Weisberg give the following advice for this situation: Using residuals to guide model development will often result in misdirection or at least more work than would otherwise necessary. To understand (2), consider the true model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$

and that x_1, x_2 and x_3 are nonlinearly related ((2) does not hold). When fitting the data without predictot x_3, the residuals plotted against x_1 show a potentially misleading nonrandom nonlinear pattern in x_1. The nonramdon patterns do not provide direct information on how the model is misspecified. Example: nyc.csv example on regression model to predict the price of dinner. To check (2), we look at the scatter plot matrix of the three continuous predictors. The predictors seem to be related linearly at least

approximately. Assuming (1) hods, we look at the standardized residuals against each predictor. Indication of the valid model!!

Another example named with data named caution in the R-library, alr3. Note the data was generated based $E(Y|X) = \frac{|x_1|}{2+(1.5+x_2)^2}$ with normal noise with mean 0 and variance 1 of 100 observations. Also note that x_1 and x_2 are close to being uncorrelated with correlation -0.043. The standardized residual vs x_2 and the fitted value indicate nonconstant error variance, which is not true. Therefore, when the (1) and (2) do not hold, we can not say anything about what part of the model is misspecified based on the residual plots.

Another example for which (2) does not hold. $Y = x_1 + x_2^2 + e$ where $E(x_2|x_1) = sin(x_1)$, x_1 is equally spaced from -3 and 3 and the errors are normally distributed with standard deviation equal to 0.1. The data can be found on the book website in the file called nonlinear.txt.The plot of standardized residual against x_1 shows periodic feature but because (2) does not hold, we can not say anything about what part of the model is misspecified.

Added variable plots

$$Y = X\beta + e \tag{3}$$

Consider the introduction of a new predictor, $Y = X\beta + Z\alpha + e$ and

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{bmatrix} \tag{4}$$

in particular, we are interested in α, the regression coefficient measuring the effect of Z on Y, having adjusted for the effect of X on Y. The added-variable plot for predictor Z enables us to visually estimate α. The added-variable plot is obtained by plotting on the vertical axis the residuals from model (3) against on the horizontal axis the residuals from

$$Z = X\delta + e \tag{5}$$

Thus the added-variable plot for predictor Z shows that part of Y that is not predicted by X against the part of Z that is not predicted by X (i.e. the effect due to X are

removed from both axes). The vector of residuals from (3), $\hat{e}_{Y.X} = (I - H_X)Y$. Note $Y = X\beta + Z\alpha + e$. Therefore, $(I-H_X)Y = (I-H_X)X\beta + (I-H_X)Z\alpha + (I-H_X)e$. Note $(I - H_X)Y$ is the residual of model (3). Consider the right hand side, $\hat{e}_{Z.X} = Z - \hat{Z} = (I - H_X)Z$. The added plot $\hat{e}_{Y.X} = \hat{e}_{Z.X}\alpha + (I - H_X)e$. It can be shown that the least square estimate of α in this model is equal to the least square estimate of α in model $Y = X\beta + Z\alpha + e$. Assuming the model is valid for the data, then the added-variable plot should produce points randomly scattered around a line through the origin with slope $\hat{\alpha}$. This plot will enable the user to identify any data points which have undue influence on the least square estimate of α.

Example on NYC data. Issue with the scatterplot matrix, it only looks at the effect of a given predictor on the response, ignoring the effects of the other predictors on the response. The shortcoming is overcome by added-variable plots. Having adjusted for the effects of the other predictors, the variable service adds little to the prediction of Y. Two points are identified as having large influence on the least squares estimate of the regression coefficient for food.

Transformations 1. Transforming only the response variable Y using inverse regression

$$g^{-1}(Y) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + e$$

Based on the results of Li and Duan (1989), Cook and Weisberg (1994) showed that if conditions (1) and (2) hold then g^{-1} can be estimated from the scatter plot of Y (on the horizontal axis)and fitted $\hat{y}$ (on the vertical axis). Such a plot is commonly referred to as inverse response plot (since the usual axis for Y is the vertical axis). It can be shown that the assumption that the predictors $X_1, X_2, \ldots, X_p$ have a multivariate normal distribution is much stronger than the assumption than the predictors $X_1, X_2, \ldots, X_p$ are linearly related at least approximately.

Example: modelling defective rates. Standard residual vs each predictor shows a curved rather than random pattern, which indicates invalid model.

$$y^{0.44} = \beta_0 + \beta_1 x_1 + \beta_{2x2} + \beta_3 x_3$$

or

$$y^{0.5} = \beta_0 + \beta_1 x_1 + \beta_{2x2} + \beta_3 x_3$$

Transform only the response using Box-Cox method

Goal is to transform the response as close to normal as possible. Example with data defects.txt. The value that maximize the log-likelihood is 0.45. In this case both inverse response plot and Box-Cox transformation method point to using square root transformation of the response. The diagnostic plots provided by R shown further confirm that square root transformation of the response is a valid model. The variable Rate is not statistically significant and the coefficient of Density is statistically significantly less than zero, The lack of statistical significance of the predictor Rate is evident in the added-variable plot.

Transforming both the response and the predictor variables

Approach 1 with two steps:

1) Transform the predictors so that the transformed predictors are as jointly normal as possible.

2. Use Inverse plot to decide on the transformation of the response.

Approach 2 :

Transform the predictors and the response simultaneously to joint normality using Box-Cox method. Example: Magazine Revenue

A scatter plot matrix of the response variable and the three predictor variables. The response variable and the three predictor variables are each highly skewed. In addition, the predictors do not appear to be linearly related. Thus, we need to consider transformations of the response and the three predictor variables. Natural logarithm of the predictors. Best $\lambda = 0.23$ from the inverse response plot. Note logarithm transformation of the response seems to be a good fit too. The variables (predictors and response have the same units and the same transformation) From the second approach, transform both predictors and response using the natural logarithm. Diagnostic plots confirm that the model fit the data well. The dashed line in the diagnostic plot (standardized residuals vs leverage) is the usual cutoff for declaring a point of high leverage. Thus observation 199 requires further investigation. Added-varaible plot indicates the lack of statistical significance of the predictor log(NewsRevenue) is evident.

Measuring multicolinearity in the regression model

Start with data bridge.txt. Note that the response variable and a number of the predictor variables are highly skewed. Thus, we need to consider transformations of the response and the five predictor variables. Notice that while the overall F-test is highly statistically significant (i.e., has a very small p-value), only one of the estimated regression coefficients is statistically significant (i.e., log(Dwgs) with a p-value <0.001). Even more troubling is the fact that the estimated regression coefficients for log(DArea) and log(Length) are of the wrong sign (i.e., negative), since longer bridges or bridges with larger area should take a longer rather than a shorter time to design. The lack of statistical significance of the predictor variables other than log(Dwgs) is evident. Note we have highly correlated predictor variables in this analysis. Variance inflation factors

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

It can be shown that

$$\mathrm{V}(\hat{\beta}_j) = \frac{1}{1 - r_{1,2}^2} \frac{\sigma^2}{(n-1)S_{x_j}^2}, j = 1, 2$$

The term $\frac{1}{1-r_{1,2}^2}$ is called a variance inflation factor. In the general multiple linear regression model,

$$\mathrm{V}(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \frac{\sigma^2}{(n-1)S_{x_j}^2}$$

and $\frac{1}{1-R_j^2}$ is called the jth variance inflation factor where R_j^2 denoted the value of multiple R^2, obtained from the regression of x_j on the other predictors.

Pitfalls of observational studies due to omitted variables

Spurious correlation: the observed association between two variables may be because both are related to a third variable that has been omitted from the regression model. Pearson studied measurements of a large collection of the skulls from the Paris catacombs. For each skull the length, breadth were measured and the correlation coefficient was computed. The correlation turns out to be significantly greater than zero. However, the discovery was deflated by his noticing that if the skulls were divided into male and female, the correlation disappeared. Pearson recognized the general nature of this phenomenon and

brought it to the attention of the world. When two measurements are correlated, this may be because they are both related to a third factor that has been omitted from the analysis. In Pearon's case, skull length and breadth were essentially uncorrelated if the factor gender were incorporated in the analysis. An example by Neyman based on fictious data which dramatically illustrates spurious correlation. The interest of the study is to empirically examine the theory of that storks brings babies and the data collected consists of the number of women, babies born and storks in each of 50 counties named as storks.txt. When regress number of babies on the number of storks, we see a strong positive association. Follow the same idea we see strong postive association between number of babies, women and storks. However, when considering the model with both number of storks and women, the estimated coefficient for storks is about 0. Thus the correlation between and the number of babies and storks are spurious as it is due to both variables being associated with the number of women. The number of women here is called omitted variable or confounding covariate.