Regression Model Specification in R/Splus and Model Diagnostics

By

Daniel B. Carr

Note 1: See 10 for a summary of diagnostics
2: Books have been written on model diagnostics. These discuss diagnostics in depth, indicate multiple criteria for looking closer for the same diagnostic, and provide guidance for corrective action. The brief description here only indicates a single criterion for a diagnostic or a short verbal indication of what to look for. The code that flags cases in R may be more sophisticated.

Note 3: A summary appears in 8.
Note 4: A few references appear in 9:

1. R Model Notation

~ means is modeled by. The variable being modeled is on the left and often referred to as the dependent variable

Example z ~ x + y

2. Specifying the X matrix for y = Xb+e.

In this notation y is the dependent variable and the columns of X are called explanatory or predictor variables. The matrix X is sometimes called the model or design matrix.

Analysts can build a model matrix X by hand. However, decades of experience has led to a short hand notation for specifying to the computer how to construct the matrix.

2.1 General

-1 means omit the default column of 1's from the models.
a + b means include predictor variables a and b as separate columns
a:b means interaction, add a column defined by ab
a*b means a + b + a:b
a %in% b means a nested within b
(a + b)^2 means a + b + a^2 + a*b + b^2

This notation extends to more than two factors. For example, a*b*c indicates a fully crossed three-way model with all the interaction terms. Similar (a+b+c)^2 include all the linear, interaction and quadratic terms

2.2 Polynomials and Splines

poly(x,3) means: x + x^2 + x^3 orthogonalized. a constant is presumed
bs() means B-Splines basis
ns() means natural cubic splines

Notes on Splines:

Splines partition the range of the predictor with k internal knot points, and used a polynomial of degree d over each interval. The d-1 derivatives are be to continuous at the k internal knot points.

An efficient modeling approach uses k+d basis functions.

Arguments:
knots: a vector of internal knot points that partition the range of the predictor
The degree of continuity can be dropped at the knot points by duplicating knot points.
df : use k = df - degree equally spaced internal knot points. Here df is the degrees of freedom for the model, not for the residuals.
degree: degree of the splines, default =3
intercept: default=F and bs() produces a matrix whose columns are orthogonal to the column of 1's

Consider basis splines versus natural cubic splines. Natural cubic splines are constrained to be linear beyond the boundary knot points (the endpoints of the data). This often improves the modeling near the endpoints. The linear constraints remove two degrees of freedom from the model.

2.3 Logical Variables and Two-Level Factors

A logical variable is equivalent to a factor with two levels. Conceptually factors produce indicator columns.

Examples
Age > 30
Sex

The idea behind the construction of model matrix for factors is easy. Consider the model y~ Age > 30. There is an implicit column of 1's to fit the mean, that conceptual matrix is as below.

		X	
Case	Mean	Age > 30	Age <= 30
1	1	1	0
2	1	1	0
3	1	0	1
...			

Conceptually there would be three coefficients to estimate for the X matrix above, one for the mean, one for Age >30 and one for Age < 30. However, the columns of X are not linearly

independent. In particular the Mean column = I(Age>30) + I(Age<=30) were I() is the indicator functions. This means the least squares solution to the problem conceptually involves a generalized inverse and that the coefficients for the three terms are not unique.

The actual model matrix constructed for in regression involves reparameterization so the three columns above becomes two, a mean column and say an increment if Age > 30. The selection of contrasts governs the reparameterization.

In general factors add a matrix of columns to X. The matrix added is reparameterized to have 1 less column than the number of factor levels and to be linearly independent of the column of 1's.

2.4 Factors with three or more levels and contrasts

If a factor has L levels, it conceptually adds L indicator columns to the model matrix. One might think of a factor variable as simpler that a continuous variable, but in the regression context a continuous variable only adds one column to the model matrix and one element to the list of parameters to be estimated.

In practice the indicator matrix for a factor and the mean column are linearly dependent. The standard approach reparameterized using contrasts. A factor with L levels adds L-1 columns to model matrix, increases the model degree of freedom by L-1 and decreases the residual degrees of freedom by the same number, L-1.

See the discussion of contrasts in for example Statistical Models in S for further information on contrasts.

3. Fitting, Residuals and Diagnostics

3.1 Conceptual description versus numerical methods

The conceptual description of the fitting process involves matrix operations including the matrix inverse operator. This is fine for understanding but poor for computation. The world of computation uses superior matrix decomposition methods such as the QR algorithm to obtain the same goals. This class will not cover QR tools in R or Splus, but they are available and used.

3.2 Fitted Values and the Projection Matrix

Assume the X matrix with n rows and p columns is of full rank, p**.** The fitted values are

$$\hat{Y} = X (X^T X)^{-1} X^T Y \qquad (1)$$

or

$$\hat{Y} = HY$$

$$\text{where} \qquad H = X\,(X^T X)^{-1}\,X^T$$

is an n x n projection matrix. H projects Y into the column space of X. This is the model space. Re-projecting makes no difference.

$$\hat{Y} = HY = HHY$$

In other words HH=H. Verify this in terms of X using (1). The name for this property is idempotent. H is idempotent and symmetric. H is called the hat matrix since it produces the estimated values denoted with the hat.

The matrix (I-H) projects y into the residual spaces. This matrix is also idempotent and symmetric.

The projection matrix H contains useful quantities. Let h_{ij} be the values of H. The sum of diagonal elements h_{ii} is the rank of X. Since X is an n x p matrix. The rank is p (unless n < p or columns are linearly dependent). Since H is an n x n matrix there are n diagonal values. Thus the average value is p/n. (Computational note: the diagonal elements of H can obtain from a singular value decomposition of X.)

When a diagonal element, h_{ii} is close to 1 the other values along i^{th} row of H are close to zero. The linear combination of observed values for y, emphasizes only the i^{th} value. As a consequence the observed and predicted values are nearly the same for the i^{th} case. In other words, the structure of the X matrix producing H, is such that observed value almost exclusively determines the predicted value. Such cases are said to have a high leverage.

High leverage cases are dangerous. If the y values are poorly measured for high leverage cases or random variation yields a unlucky observation, the fit to the remaining values can be badly distorted.

In general, cases with leverage above 2*p/n are worth further assessment. If value seems suspect, it usually pays to see if scientific criteria provide as basis for removing or down weighting that case.

Least squares fit high-leverage points. If the dependent values for such cases are anomalous least squares will be silent. The model will fit them. One often needs to look at the residuals from low-leverage points to get clues that something might be wrong.

3.3 Residuals

The residuals are given by

$$e = (I-H)\,Y$$

The residuals, e, are estimates of the errors in conceptual the model. The hat has been omitted for convenience.

Like H, I-H is also symmetric and idempotent. It projects the dependent values into the residual space.

In the classic model, the theoretical errors have a standard deviation of sigma and are independent identically distributed.

A check on this assumption is **Diagnostic Plot 1: The spread location plot**. The square root of absolute residuals in plotted on the y-axis versus the predicted values on the x-axis. If a wedge shape appears the residuals have a variance that is related to the dependent variables. A transformation of the dependent variables may address this problem or modeling based on another family of distributions can address the problem. If the dependent variable follows a Poisson distribution the variance does increase with the mean.

The basic model is

$$Y \sim N(XB, \sigma^2 I) \qquad T$$

Thus $e = (I-H)Y \sim N((I-H)XB, \sigma^2 (I-H)I(I-H))$

Simplifying $e \sim N(0, \sigma^2 (I-H))$

We can use the above to produce standardized or studentized residuals. The theoretical standard error of the i^{th} residual under the standard normal model above has mean zero and standard deviation $s * \sqrt{(1-h_{ii})}$ where s is an estimate of sigma. Dividing the residuals by the estimated standard deviations yields standard residuals.

An alternative approximation for sigma, denoted s_{-i}, leaves out the i^{th} case in the computation. Dividing the residuals by the leave-one-out the approximation for sigma yields studentized residuals. (Only one regression is required, since the leave one out estimates case can be handled algebraically from all case regression results.) Conceptually the studentized residual is more like a student-t statistic than a standardized residual since for the typical t-statistic construction, the numerator and the denominator are supposed to be independent.

The lead to our next two diagnostics plots.

Diagnostics Plot 2: A normal QQplot using studentized residuals.

Diagnostics Plot 3: A plot of Studentized residual versus case number

We look for thick tails in plot 2 and for values bigger that ±2 in plot 3.

In practice is the choice between standardized and studentized residuals doesn't usually make much difference unless n is small. Also, that fact the there is a small correlation among the residuals is usually ignored in the QQplot with little harm done.

Remember that high leverage cases that may have small residuals.

4. Leverage and Influence Versus Case plots

4.1 Leverage

The diagonal elements of the hat matrix show in section 3.2 indicated the leverage of case in the regression model. This lead to

Diagnostics Plot 4: A plot of hat values h_{ii} versus case number

Cases with values larger 2p/n have high leverage and should draw attention

4.2 Leaving One Case Out Influence

One way to assess the influence of a case is by assessing change in regression model due to leaving the case out. Different statistics can show different facets of influence. Sometimes the influence is spread out and global and sometime influence is localized . Cook's distance D_i assesses the influence on the model as a whole.

$D_i = z_i^2 h_{ii} / (p(1-h_{ii}))$. This leads to

Diagnostics Plot 5: Cook's distance versus case number

Various criteria can be used to assess the values. Usually influential cases have $D_i > 4n$

Each of the regression coefficients can change from leaving out a case. This leads to DFBetas on For each term in the model include the constant term. The formula for the j^{th} term is

$DFBetas_j = (b_j - b_{j,-i})/(s_{-i}/sqrt(RSS_j))$

Here the -1 subscript indicates the statistic is estimated by leaving out the i^{th} case and RSS_j is the residual sum of squares from regressing the j^{th} predictor on all the other predictors.

Diagnostics Plots 6: DFBetas versus case number

Again various criteria can be used to assess the values. On criterion look more closely at cases with absolution values > 2/sqrt(n).

The statistis $DFFits_i$ assess the change in the predicted value for the i^{th} case from leaving the i^{th} case in building the model and the using the ith case predictors in this model.

$DFFits_i = (Y_i - Y_{i,-i})/(s_{-i}/sqrt(h_{ii}))$

Diagnostics Plot 7 : DFFits versus case number

5. Simple lack of Fit

Plot the residuals versus each predictor variable

Diagnostics Plots 8 : Residuals versus Predictors

6. Partial Regression Plots

Partial residual plots, one for each variable, regress each predictor j on all the other predictors and use the residual to indicate in unique contribution to the j^{th} predictor to the model space. This is shown on the x axis. The partial residual plot construction also determines how much of Y has not been modeled by all the other predictors. That it uses the residual from regression Y on all the other predictors to plot on the y-axis. The slope of the best fit line for the j^{th} plot is b_j the

coefficient from the full regression models. However this plot shows if the slope is being determined by one or just a few points. We may not want to include a variable in the model if its only merit is to fit one or just a few cases. I have examples where the statistical significant of a coefficient in a model was due to helping the fit of a case

Detrended partial residual plots are similar except the y axis variable is the residual of fitting Y to the full model. In both cases we look for the influence of a few points or for nonlinearity.

Diagnostic Plots 9 : Partial Residual (or Added Variable) plots versus Predictors
Diagnostic Plots 10: Detrended Partial Residual Plots

7. Durban Watson Statistic for Correlation

Data collect over times exhibits temporal dependence or correlational structure often enough that is worth taking a look. The Durban Watson statistic is the classic statistic for seeing if there is lag 1 correlation in the model residuals. The statistic is the ratio of the sum of differences squared for time adjacent residuals divided by the sum of squares of residual. There are is a table for critical values. This is a useful statistic but not stressed in the class on visualization.

8. Diagnostics Plot Summary

1. Spread Location Plot
   x= predicted values
   y=(absolute residuals)^.5
   Reference: horizontal loess line
   Look for: wedge shapes that imply lack of homogeneity
   Look for: curves that imply non-linearity

2. QQ plot
   x=standard normal quantiles
   y=sorted residuals
   Reference: robust straight line
   Look for: departures from normality
   Thick-tails suggest problems such as contamination

3. Outlier plot
   x=Observation No.
   y=Studentized residuals
   Reference: horizontal lines at + 2 and -2
   Look for: outliers

4. Leverage points
   x=Observation No.
   y=Influence
   Reference: horizontal line at 2p/n
   Look for: highly influential points

5. Influential points for the model as whole
   x = Observation No.
   y = Cook's Distance
   Examine cases with values > 4/ n

6. Influential points for individual coefficients
   x= Observation No.
   y= $DFBeta_j$
   Examine cases with absolute values > 2/sqrt(n)

7. Influential points for predicted values
   x = Observation No.
   y = DFFits
   Examine cases with absolute values > 2sqrt(p/n)

8. Simple Lack of Fit (for each variables)
   x=independent variable
   y= residuals
   Reference: horizontal line versus loess line
   Look for: nonlinearity

9. Partial residual plots (for each variable)
   x = residual of Xi on X-i
   y = residual of Y on X-i
   Reference: line for fit of y on x
   Look for: fitting just a few points or nonlinearity

10. Detrended partial residual plots Added Variable Plot (for each variable)
    x = residual of Xi on X-i
    y = residual of Y on X
    Reference: horizontal line versus loess line
    Look for: fitting just a few points or nonlinearity

9. References

Chamber, J. M and T. J. Hastie, Eds: 1992. *Statistical Models in S.* Wadsworth & Brooks/Cole Pacific Grove, CA.

Hamilton, Lawrence C. 1992. *Regression with Graphics,* Brook/Cole Publishing Company Pacific Grove California.

Cook, R. D. and S. Weisberg. 1982. *Residuals and Influence in Regression.* Chapman and Hall, New York.

Besley, D. A, E. Kuh, and R. E. Welsch. 1980. *Regression Diagnosis,* Wiley, New York.