

## Stat 328 Handout on Diagnostics

There are various kinds of residuals, ways of plotting them and measures of "influence" on a regression that are meant to help in the black art of model building. We have already alluded to the fact that under the MLR model, we expect **ordinary residuals**

$$e_i = y_i - \hat{y}_i$$

to look like mean 0 normal random noise and that **standardized or studentized residuals**

$$e_i^* = \frac{e_i}{\text{standard error of } e_i}$$

should like standard normal random noise.

### Deleted Residuals and the PRESS Statistic

There is also the notion of **deleted residuals**. These are built on the idea that a model should not be terribly sensitive to individual data points used to fit it, or equivalently that one ought to be able to predict a response even without using that response to fit the model. Beginning with a particular form for a MLR model and  $n$  data points, let

$$\hat{y}_{(i)} = \text{the value of } y_i \text{ predicted by a model fit to the other } (n - 1) \text{ data points}$$

(note that this is not necessarily  $\hat{y}_i$ ). The  $i$ th deleted residual is

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

and the hope that if a model is a good one and not overly sensitive to the exact data vectors used to fit it, these shouldn't be ridiculously larger in magnitude than the regular residuals,  $e_i$ . The "**prediction sum of squares**" is a single number summary of these

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_{(i)})^2$$

and one wants small values of this. (Note that  $PRESS \geq SSE$ , but one hopes that it is not too much larger.)

This does not exhaust the ways in which people have suggested using the residual idea. It is possible to invent **standardized/Studentized deleted residuals**

$$e_{(i)}^* = \frac{e_{(i)}}{\text{standard error of } e_{(i)}}$$

and there are yet other possibilities.

## Partial Residual Plots (JMP "Effect Leverage Plots")

In somewhat nonstandard language, SAS/JMP makes what it calls "effect leverage plots" that accompany its "effect tests." These are based on another kind of residuals, sometimes called **partial residuals**. With  $k$  predictor variables, I might think about understanding the importance of variable  $j$  by considering residuals computed **using only the other  $k - 1$  predictor variables to do prediction (i.e. using a reduced model not including  $x_j$ )**. Although it is nearly impossible to see this from their manual and help functions or how the axes of the plots are labeled, the effect leverage plot in JMP for variable  $j$  is essentially a plot of

$$e^{(j)}(y_i) = \text{the } i\text{th } y \text{ residual regressing on all predictor variables except } x_j$$

versus

$$e^{(j)}(x_{ji}) = \text{the } i\text{th } x_j \text{ residual regressing on all predictor variables except } x_j$$

To be more precise, exactly what is plotted is

$$e^{(j)}(y_i) + \bar{y} \text{ versus } e^{(j)}(x_{ji}) + \bar{x}_j$$

On this plots there is a horizontal line drawn at  $\bar{y}$  (**at  $y$  partial residual equal to 0**, i.e.  $y$  perfectly predicted by all predictors excepting  $x_j$ ). The vertical axis IS in the original  $y$  units, but should not really be labeled as  $y$ , but rather as partial residual. The sum of squared vertical distances from the plotted points to this line is then  $SSE$  for a model without predictor  $j$ .

The horizontal plotting positions of the points are in the original  $x_j$  units, but are essentially partial residuals of the  $x_j$ 's NOT  $x_j$ 's themselves. The horizontal center of the plot is at  $\bar{x}_j$  (**at  $x_j$  partial residual equal to 0**, i.e. at  $x_j$  perfectly predicted from all predictors except  $x_j$ ). The non-horizontal line on the plots is in fact the least squares line through the plotted points. What is interesting is that the usual residuals from that least squares line are the residuals for the full MLR fit to the data. So the sum of the squared vertical distances from points to sloped line is then  $SSE$  for the full model. The larger is reduction in  $SSE$  from the horizontal line to the sloped one, the smaller the  $p$ -value for testing  $H_0: \beta_j = 0$ .

Highlighting a point on a JMP partial residual plot makes it bigger on the other plots and highlights it in the data table (for examination or, for example, potential exclusion). We can at least on these plots see which points are fit poorly in a model that excludes a given predictor and the effect the addition of that last predictor has on the prediction of that  $y$ . (Note that points near the center of the horizontal scale are ones that have  $x_j$  that can already be predicted from the other  $x$ 's and so addition of  $x_j$  to the prediction equation does not much change the residual. Points far to the right or left of center have values of predictor  $j$  that are unlike their predictions from the other  $x$ 's. They both tend to more strongly influence the nature of the change in the model predictions as  $x_j$  is added to the model, and tend to have their residuals more strongly affected than points in the middle of the plot (where  $x_j$  might be predicted from the other  $x$ 's).

## Leverage

The notion of how much potential influence a single data point has on a fit is an important one. The JMP partial residual plot/"effect leverage" plot is aimed at addressing this issue by highlighting points with large  $x_j$  partial residuals. Another notion of the same kind is based on the fact that there are  $n^2$  numbers  $h_{ii'}$  ( $i = 1, \dots, n$  and  $i' = 1, \dots, n$ ) depending upon the  $n$  vectors  $(x_{1i}, x_{2i}, \dots, x_{ki})$  only (and not the  $y$ 's) so that each  $\hat{y}_i$  is

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \dots + h_{i,i-1}y_{i-1} + h_{ii}y_i + h_{i,i+1}y_{i+1} + \dots + h_{in}y_n$$

$h_{ii}$  is then somehow a measure of how heavily  $y_i$  is counted in its own prediction and is usually called **the leverage** corresponding data point. (JMP calls the of the  $h_{ii}$  the "hats.") It is a fact that  $0 < h_{ii} < 1$  and  $\sum_{i=1}^n h_{ii} = k + 1$ . So the  $h_{ii}$ 's average to  $(k + 1)/n$ , and a plausible rule of thumb is that when a single  $h_{ii}$  is more than twice this average value, the corresponding data point has an important  $(x_{1i}, x_{2i}, \dots, x_{ki})$ .

It is not at all obvious, but as it turns out, the  $i$ th deleted residual is  $e_{(i)} = e_i/(1 - h_{ii})$  and the *PRESS* statistic has the formula  $PRESS = \sum_{i=1}^n \left( \frac{e_i}{1-h_{ii}} \right)^2$  involving these leverage values. This shows that big *PRESS* occurs when big leverages are associated with large ordinary residuals.

## Cook's D

The leverage  $h_{ii}$  involves only predictors and no  $y$ 's. A proposal by Cook to measure the overall effect that case  $i$  has on the regression is the statistic

$$D_i = \frac{h_{ii}}{(k + 1)MSE} \left( \frac{e_i}{1 - h_{ii}} \right)^2 = \frac{h_{ii}}{(k + 1)} \left( \frac{e_{(i)}}{s_e} \right)^2$$

(abbreviating  $\sqrt{MSE}$  as  $s_e$  as in Dielman) where large values of this identify points that by virtue of either their leverage or their large (ordinary or) deleted residual are "influential."

$D_i$  is **Cook's Distance**. The second expression for  $D_i$  is product of two ratios. The first of these is a "fraction of the overall total leverage due to  $i$ th case" and the second is an approximation to the square of a standardized version of the  $i$ th deleted residual. So  $D_i$  will be large if case  $i$  is located near the "edge" of the data set in terms of the values of the predictors, AND has a  $y$  that is poorly predicted if case  $i$  is not included in the model fitting.