Outliers. Leverage, and Influential Points in Regression

1 Introduction

1.1 Some General Ideas

The notions of "outlying" or "influential" data are deliberately somewhat vague notions in Statistics. In linear models we usually deal with three related notions:

- A data point $(y_i, x_{i1}, \ldots, x_{ik})$ is *outlying* if it is far from the remainder of the data.
- A data point has high *leverage* if parameter estimates or fitted values change a great deal when the data point is changed by a small amount—this is an inherently differential-calculus notion.
- A data point is *influential* if removing it from the data set changes the parameter estimates or fitted values a great deal.

These notions are nicely illustrated in Figure 1.

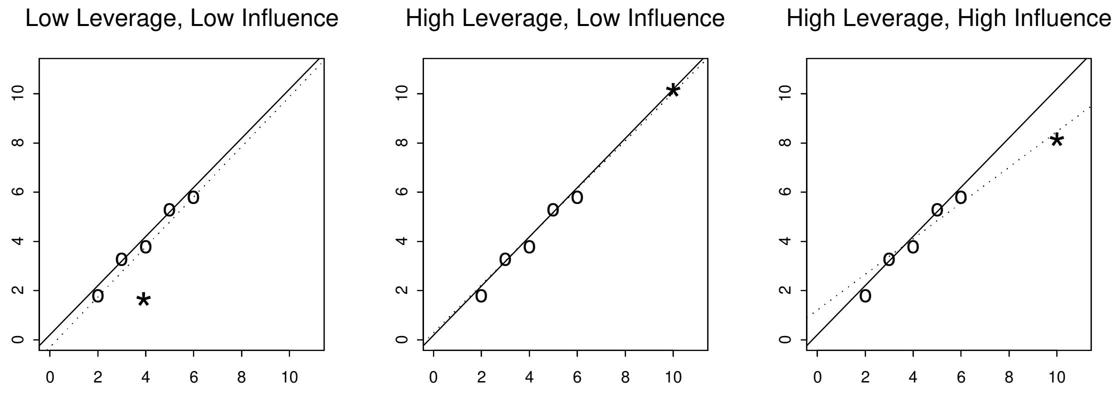


Figure 1: Outlying points ($*$) illustrating varying degrees of leverage and influence. The solid line is the linear fit without the oulier, the dotted line is the fit with the outlier.

From the figure, we can deduce sevearal important heuristics for linear regression:

- Outlyingness in X seems to drive *leverage*, but may or may not be related to lack of model fit.
- Outlyingness in Y is a clear indication of *lack of model fit*; and combined with high leverage seems to drive influence: *Influence* $=$ *Leverage* $\times$ *Outlyingness.*
- To measure *influence*, compare the regression with and without the data point.

In this handout we will explore more formally some quantitative measures of outlyingness, leverage, and influence. You should keep in mind that these are *only* rules of thumb; what (if anything) you do with a data point that is labelled as unusual by one or more of these measures is ultimately *qualitative decisions* that you have to make in the scientific context of the problem you are working on.

Indeed, *it it important to investigate why a data point is unusual*:

- If an outlier is found to be due to a coding error or a procedural error in measurement, it may be corrected and saved, or discarded; either way we have improved the data.
- If an outlier is correctly measured and coded, it represents a real departure from the model we thought was appropriate and this is important information:
  - We may wish to modify the model (add a polynomial term, transform the X and/or Y values, or change the error distribution, e.g. from a Normal to a t on a few degrees of freedom);
  - Or at the very least include special coding for points similar to this one.

1.2 Relating Regression Equations—"Leave-one-out" Diagnostics

The basic idea of many least-squares-based diagnostic tools for outliers, leverage, and especially influence, are to examine the regression with and without the i^{th} data point. To check each of n data points in turn requires refitting the regression equation $n+1$ times. This could be computationally intensive, but we are saved in least-squares regression by a remarkable formula that relates the fitted regression with and without the i^{th} case.

In the regression model

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

suppose that X is $n \times (k+1)$, Y is $n \times 1$, and let $X_{(-i)}$ be the $(n-1) \times (k+1)$ matrix with the i^{th} row deleted, and similarly for $\underline{y}_{(-i)}$.

Suppose and the matrix $(X^TX)^{-1}$ has been computed; then Weisberg (1985, *Applied Linear Regression*, Wiley, p. 293) gives the following formula, to calculate $(X^T_{(-i)}X_{(-i)})^{-1}$ from $(X^TX)^{-1}$:

$$(X^T_{(-i)}X_{(-i)})^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}\underline{x}_i\underline{x}_i^T(X^TX)^{-1}}{1-h_{ii}} \tag{1}$$

where $\underline{x}_i$ is the i^{th} row of X, written as a column vector (this is the same convention we have used before), and h_{ii} is the i^{th} diagonal element of the hat matrix $H = X(X^TX)^{-1}X^T$. Gauss used a version of equation (1) as far back as 1821; a history of it and many variations is given by Henderson and Searle (1981, *SIAM Review, 23,* pp. 53–60). It can be verified, with some effort, by multiplying by $X^T_{(-i)}X_{(-i)}$ on either side and showing that one obtains the $(k+1) \times (k+1)$ identity matrix.

Equation (1) can be applied to give all the results that one would want relating the regressions with and without the i^{th} case. For example, since $X^T_{(-i)}\underline{y}_{(-i)} = X^T\underline{y} - \underline{x}_i \cdot y_i$, one can use equation (1) to relate the regression coefficients $\underline{\hat{\beta}}_{(-i)}$, estimated without the i^{th} case, to the coefficients $\underline{\hat{\beta}}$ estimated with all the data:

$$\underline{\hat{\beta}}_{(-i)} = (X^T_{(-i)}X_{(-i)})^{-1}X^T_{(-i)}\underline{y}_{(-i)} = \underline{\hat{\beta}} - \frac{(X^TX)^{-1}\underline{x}_i(y_i - \hat{y}_i)}{1-h_{ii}} = \underline{\hat{\beta}} - \frac{(X^TX)^{-1}\underline{x}_iE_i}{1-h_{ii}} \tag{2}$$

Equation (2), and other facts derived from equation (1), are used repeatedly in the assessment of unusual data in linear regression. These and related identities are developed in detail in appendices A and B of Myers (1990, *Classical and Modern Regression*, Duxbury Press).

2 Outliers in Y

Outlyingness is often measured with respect to a reference distribution such as the Normal distribution. For example the "boxplot rule," that a point is an outlier if it is more than 1.5 IQR's from the nearest quartile, is based on the fact that for the normal distribution, 1.5IQR's from the nearest quartile is the same as median $\pm$ 2IQR's, $\approx$ mean $\pm$ 3SD's: we are basically identifying points as outliers when they lie outside a robust guess about the $\pm$ 3SD range.

In regression analysis we know that

$$\begin{aligned} \underline{\hat{y}} &= H\underline{y} \qquad \text{and} \\ \mathrm{Var}\,(\underline{\hat{y}}) &= \sigma_E^2 H, \end{aligned}$$

where $H = X(X^TX)^{-1}X^T$ is the hat matrix, and so it is natural to inspect the *standardized residuals*

$$E_i' = \frac{E_i}{S_E\sqrt{1-h_{ii}}}$$

for outliers or violations of the "normal errors" assumption. We have used these before to check for outlying residuals (e.g., $|E_i'| > 2$ or 3) and to check for symmetry, unimodality, or normality of the residuals (e.g. with a boxplot, stem and leaf plot, or `qqnorm` plot).

The E_i''s have two problems: first, they are dependent, and second, even though the standardization suggests they should be normally distributed when the full regression model holds, their exact distribution is difficult to calculate. One way around this is to make their numerator and denominator independent of one another. This is what the *studentized residuals* do:

$$E_i^* = \frac{E_i}{S_{E(-i)}\sqrt{1-h_{ii}}} \tag{3}$$

where $S_{E(-i)}$ is calculated from the regression with the i^{th} case deleted. These are not difficult to calculate, since it follows from (1) that

$$S_{E(-i)}^2 = S_E^2 \cdot \frac{n-k-1-(E_i')^2}{n-k-2} \tag{4}$$

Because they are based on different parts of the data, the numerator and denominator of equation (3) are independent, and it is not hard to show that each E_i^* follows a t distribution on $n-k-2$ degrees of freedom.

Example. Studentized Residuals.

We can easily build from our earlier function `stdres` to calculate standardized residuals so that it calculates studentized residuals. The earlier function was

```
stdres <- function(fitted.lm)
{
        resids <- fitted.lm$residuals
        df <- fitted.lm$df.residual
        S2E <- sum(resids^2)/df
        hats <- lm.hat(fitted.lm)
        stdres <- resids/sqrt((1 - hats) * S2E)
        return(stdres)
}
```

and we use this together with equation (4) to get the studentized residuals

```
studres <- function(fitted.lm)
{
        zres <- stdres(fitted.lm)
        df <- fitted.lm$df.residual # n-k-1
        tres <- zres * sqrt((df-1)/(df-zres^2))
        return(tres)
}
```

Recall from earlier in the semester the Fuel data with variables

```
       (State) = Two-Letter State Code
         Pop   = Population in each state
         Tax   = 1971 Gas tax, cents per gallon
         Nlic  = 1971 thousands of drivers
         Inc   = 1972 per capita income
         Road  = 1971 miles of primary highways
         Fuelc = fuel consumption, millions of gallons
         Dlic  = 100*Nlic/Pop = proportion of licensed drivers
         Fuel  = 1000*Fuelc/Pop = per-capita fuel consumption
```

The data set `fuel48` has data for the 48 contiguous states in the United States (we deleted Alaska and Hawaii in an earlier look at outliers and influential data points). We fitted the regression

```
> summary(mod3 _ lm(formula = Fuel~ Dlic + Tax, data = fuel48))
Call: lm(formula = Fuel ~ Dlic + Tax, data = fuel48)
Residuals:
    Min     1Q Median    3Q   Max
 -123.2 -60.17 -2.908 45.03 242.6

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  108.9709  171.7859     0.6343    0.5291
       Dlic   12.5149    2.0906     5.9862    0.0000
        Tax  -32.0753   12.1972    -2.6297    0.0117

Residual standard error: 76.13 on 45 degrees of freedom
Multiple R-Squared: 0.5567
F-statistic: 28.25 on 2 and 45 degrees of freedom, the p-value is 1.125e-08

Correlation of Coefficients:
```

```
     (Intercept)    Dlic
Dlic -0.8509
 Tax -0.7444      0.2880

> zres _ stdres(mod3)
> tres _ studres(mod3)
> cbind(Fitted=fitted(mod3),fuel48[,c("Fuel", "Dlic", "Tax")],
+ stdres=zres,studres=tres,hii=lm.hat(mod3))[tres>2,]
     Fitted Fuel Dlic Tax   stdres  studres        hii
ND 560.2461  714 54.0   7 2.065885 2.147158 0.04437353
WY 725.4423  968 67.2   7 3.345335 3.816379 0.09301652
```

As might be expected from equation (4), when n is large and the residuals are all fairly moderate, $S^2_{E(-i)} \approx S^2_E$ and so there isn't much practical difference between standardized and studentized residuals. A direct comparison in the fuel example confirms this:

```
> rbind(tres=studres(mod3),zres=stdres(mod3))[,order(studres(mod3))]
            RI        MA        CA        CO        MN        MI        OH
tres -1.672340 -1.583805 -1.383465 -1.089977 -1.063374 -1.039405 -1.035494
zres -1.639922 -1.557909 -1.369626 -1.087707 -1.061832 -1.038478 -1.034664
            NJ        NY         KS         WV         DE         WI
tres -1.030547 -1.020213 -0.8911714 -0.8799859 -0.8765339 -0.7814604
zres -1.029837 -1.019750 -0.8932164 -0.8822003 -0.8787991 -0.7848629
             IL         OR         PA         CN         NE         OK
tres -0.7255214 -0.7232718 -0.6691335 -0.6527026 -0.6155316 -0.5527864
zres -0.7293698 -0.7271349 -0.6732776 -0.6569055 -0.6198240 -0.5571016
             WN         ID         TX         NH          AZ         MS
tres -0.3350501 -0.2507156 -0.2471201 -0.1633223 -0.09373687 0.01673907
zres -0.3384047 -0.2533677 -0.2497392 -0.1651178 -0.09478661 0.01692816
             MO         MD        VT        FL        IA        LA        NV
tres 0.03572770 0.05662331 0.2000586 0.2241517 0.2269398 0.3390906 0.3408160
zres 0.03613089 0.05726105 0.2022272 0.2265553 0.2293700 0.3424751 0.3442131
            AL        GA        SC        TN        IN        ME        NC
tres 0.3709036 0.5002055 0.5122468 0.5149334 0.8566965 0.8632288 0.8782367
zres 0.3745097 0.5044255 0.5164972 0.5191900 0.8592405 0.8656834 0.8804769
            UT        AR       SD       VA       MT       KY       NM       ND
tres 0.9647025 0.9989029 1.085339 1.087456 1.154539 1.340104 1.489349 2.147158
zres 0.9654467 0.9989273 1.083199 1.085256 1.150291 1.328408 1.469591 2.065885
           WY
tres 3.816379
zres 3.345335
```

However when n is small or there are a few very substantial residuals, studentized residuals behave better. □

It also turns out that E_i^* is the t-statistic for testing the hypothesis $H_0: \ \gamma = 0$ in the regression model

$$\underline{y} = X\underline{\beta} + \gamma D + \underline{\epsilon}$$

where $D = (0, 0, \ldots, 0, 1, 0, 0, \ldots, 0)^T$ is the dummy variable with a 1 in row i and 0's in every other row.

Example. A Studentized Residual As a t-statistic For a Dummy Variable.

```
> D _ rep(0,length=48)
> names(D) _ dimnames(fuel48)[[1]]
> D["WY"] _ 1
> D
 ME NH VT MA RI CN NY NJ PA OH IN IL MI WI MN IA MO ND SD NE KS DE MD VA WV NC
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 SC GA FL KY TN AL MS AR LA OK TX MT ID WY CO NM AZ UT NV WN OR CA
  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0
> summary(mod4 _ lm(formula = Fuel~ Dlic + Tax + D, data = fuel48))

[stuff skipped]

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  198.6515  152.4054     1.3034    0.1992
       Dlic   10.6912    1.8939     5.6452    0.0000
        Tax  -30.9331   10.6959    -2.8921    0.0059
          D  267.4335   70.0752     3.8164    0.0004

[stuff skipped]
```

The t-statistic for the coefficient of D is clearly the same as the studentized residual we calculated above, for Wyoming. □

3 Leverage and outliers in X

Leverage is the extent to which small changes in a data point cause large changes in the fitted values of the regression —this is an inherently differential-calculus notion. We saw for example that since

$$\underline{\hat{y}} = X(X^T X)^{-1} X^T \underline{y} = H\underline{y}$$

it follows that

$$\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii}$$

and so h_{ii} is a sensible measure of leverage, in this differential-calculus sense.

Digression: Mahalanobis distance.

It seemed in Figure 1 that outlyingness in X was somehow related to leverage. We can verify this intuition if we measure distance in the right way.

In simple regression with an intercept we have seen that

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^n (x_j - \overline{x})^2}$$

this can be interpreted as $\frac{1}{n}$, plus the squared distance from the point x_i to the center of all of the x's, normalized by the "variance" of the x's. Similarly, one can show (Weisberg, 1985, p. 111) that in multiple regression

$$h_{ii} = \underline{x}_i^T (X^T X)^{-1} \underline{x}_i = \frac{1}{n} + (\underline{x}_i^* - \underline{\overline{x}^*})(X^{*T} X^*)^{-1} (\underline{x}_i^* - \underline{\overline{x}^*})^T$$

where $\underline{\overline{x}^*}$ is the $(n-1)$-dimensional vector of means of the columns of X, not including the intercept column (column of 1's), and where X^* is obtained from X by deleting the column of 1's (intercept) and subtracting $\underline{\overline{x}^*}$ from each row.

The quadratic form on the right is called the *Mahalanobis distance* from $\underline{x}_i$ to $\underline{\overline{x}}$; note that the Mahalanbis distance is constant on ellipsoids centered at $\underline{\overline{x}}$.

End of Digression.

Note that $\overline{h} = \text{trace}\,(H)/n = (k+1)/n$; in practice, $h_{ii} > 2\overline{h}$ or $3\overline{h}$ are sometimes used as cutoffs for "large influence". I like to use this in combination with graphical display like a stem and leaf plot or scatter plot of the hat values, to see that the "large influence" h_{ii}'s are also well-separated from the other h_{ii}'s.

Example. Picking out hat values.

The last time we looked at the Fuel data, we looked at hat matrix values as follows

```
> stem(lm.hat(mod3))

N = 48   Median = 0.0443331
Quartiles = 0.0324329, 0.0712345

Decimal point is 2 places to the left of the colon

    2 : 2345677
    3 : 112223344567789
    4 : 1446
    5 :
    6 : 12377779
    7 : 0023
    8 : 15
    9 : 3
   10 :
   11 :
   12 : 137
   13 :
   14 : 7

High: 0.1607706 0.1843257 0.2066504

> cbind(Fitted=fitted(mod3),fuel48[,c("Fuel", "Dlic", "Tax")],
+ stdres=stdres(mod3),hii=lm.hat(mod3))[lm.hat(mod3) > 0.16,]
     Fitted Fuel Dlic Tax     stdres       hii
```

```
CN 502.8162  457 57.1  10 -0.6569055 0.1607706
SD 790.5195  865 72.4   7  1.0831988 0.1843257
TX 656.9354  640 56.6   5 -0.2497392 0.2066504
```

In this case,

$$\overline{h} = \frac{3}{48} = 0.0625; \qquad 2\overline{h} = 0.1250; \qquad 3\overline{h} = 0.1875$$

so we are picking out values, based on outlyingness, similar to those we would have picked using the $\overline{h}$ rules of thumb. □

4 Influence

There are two common measures of influence in regression analysis, and many variations on these. I will only deal with

- Belsley, Kuh, and Welch's (1980) DFBETA and DFBETAS statistics; and
- Cook's (1977) D statistics.

The most direct measure of influence on the regression is to fit the regression with and without the i^{th} data point, and compare the regression coefficients. The statistics

$$DFBETA_{ij} \equiv D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)} \tag{5}$$

are easy to calculate and often very helpful in determining which data points influence which coefficients. A natural alternative is to consider

$$DFBETAS_{ij} \equiv D^*_{ij} = \frac{D_{ij}}{SE_{(-i)}(\hat{\beta}_j)}$$

and use cutoffs of $\pm$ 1, 2 or 3 to indicate influential i's for each j. However, the best approach is usually graphical: for each coefficient β_j, make a plot of all the $DFBETA_{ij}$'s, and inspect data points that produce large values.

Cook's D statistics are intended to summarize the information available in the $DFBETA$ statistics. For each i, define

$$\begin{aligned} D_i &= \frac{(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(-i)})^T (X^T X)(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(-i)})}{(k+1)S_E^2} \\ &= \frac{(\underline{\hat{y}} - \underline{\hat{y}}_{(-i)})^T (\underline{\hat{y}} - \underline{\hat{y}}_{(-i)})}{(k+1)S_E^2} \end{aligned} \tag{6}$$

The first line has the form of an F-statistic for "testing" whether $D_{ij} = 0 \ \forall \ j$; the distribution theory does not go through, but D_i is still sensitive to differences between $\underline{\hat{\beta}}$ and $\underline{\hat{\beta}}_{(-i)}$.

The second line follows since $\underline{\hat{y}} = X\underline{\hat{\beta}}$ (and similarly for $\underline{\hat{\beta}}_{(-i)}$). Clearly, D_i is also a measure of discrepancy between the fitted values for $\underline{y}$ under regression including the i^{th} case, vs. excluding the i^{th} case. Using equation (2), we can easily show that

$$\begin{aligned} D_i &= \frac{(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(-i)})^T (X^T X)(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(-i)})}{(k+1)S_E^2} \\ &= \frac{[(X^T X)^{-1}\underline{x}_i E_i]^T (X^T X)[(X^T X)^{-1}\underline{x}_i E_i]}{(k+1)S_E^2(1-h_{ii})} \\ &= \frac{E_i^2 \underline{x}_i^T (X^T X)^{-1}\underline{x}_i}{(k+1)S_E^2(1-h_{ii})} \\ &= \frac{(E_i')^2}{k+1} \cdot \frac{h_{ii}}{1-h_{ii}} \end{aligned} \tag{7}$$

The latter equation shows that Cook's D is simultaneously a measure of outlyingness ($(E_i')^2$ large) and leverage (h_{ii} large).

Cutoffs for "large" D_i exist, but it is generally better to plot the D_i, identify a few large values and inspect those data points directly.

Example. Cook's D and DFBETA.

Probably a reasonable approach in general is to plot Cook's D's first to get an overall idea of whether there are any influential points, and then do scatter plots of the individual D_{ij}'s for all i and each fixed j to find out which data points are influencing which coefficients.

The standard diagnostic plots in Splus give Cook's D directly; we have shown them again for the Fuel data in Figure 2. The three data points identified by Cook's D are 18, 19, and 40. Using equation (7) it is easy to identify the three largest D values as well:

```
> plot(mod3 _ lm(formula = Fuel ~ Dlic + Tax, data = fuel48))
> zres _ stdres(mod3)
> hats _ lm.hat(mod3)
> cooks _ zres^2/3 * hats/(1-hats)
> names(cooks) _ dimnames(fuel48)[[1]]
> sort(cooks)[46:48]
         ND        SD        WY
 0.06605824 0.0883821 0.3825769
> order(cooks)[46:48]
[1] 18 19 40
```

We first re-check the residual and h_{ii} diagnostics for these points, and then produce "index plots" (time series plots) for each DFBETA.

```
> zres _ stdres(mod3)
> tres _ studres(mod3)
> cbind(Fitted=fitted(mod3),fuel48[,c("Fuel", "Dlic", "Tax")],
+ stdres=zres,studres=tres,hii=lm.hat(mod3))[c(18,19,40),]
     Fitted Fuel Dlic Tax   stdres  studres        hii
ND 560.2461  714 54.0   7 2.065885 2.147158 0.04437353
SD 790.5195  865 72.4   7 1.083199 1.085339 0.18432572
```

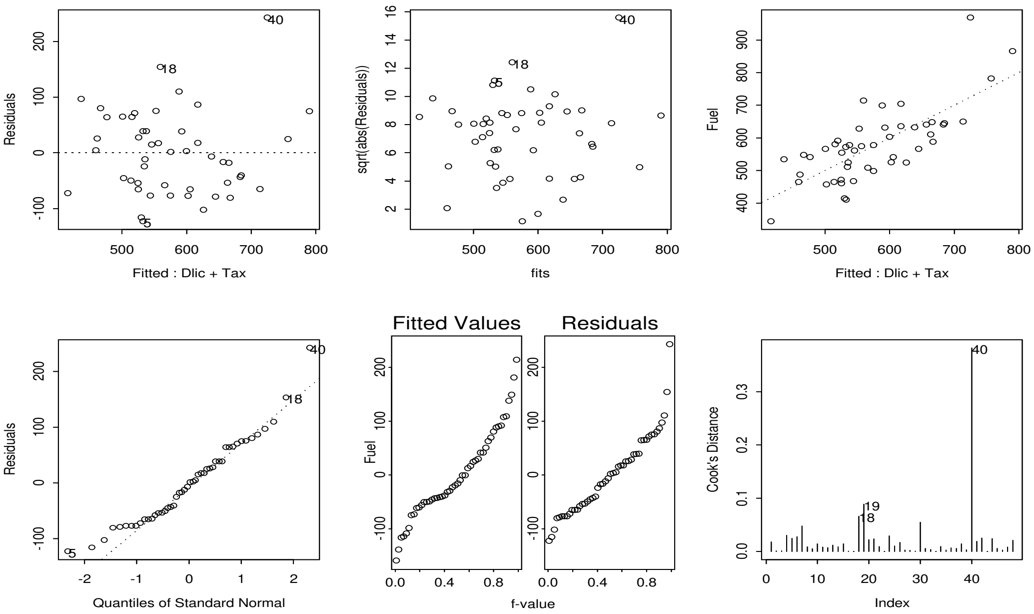


Figure 2: Splus's standard diagnistic plots for `lm(formula = Fuel ~ Dlic + Tax, data = fuel48)`.

```
WY 725.4423  968 67.2   7 3.345335 3.816379 0.09301652
> X _ model.matrix(mod3)
> resids _ residuals(mod3)
> hats _ lm.hat(mod3)
> DFBETA _ solve(t(X)%*%X) %*% t(X) %*% diag(resids/(1-hats))
> # (see note following this SPLUS code extract)
> for (j in 1:3) {
+ plot(DFBETA[j,], ylab=paste("beta",j-1,sep=""))
+ identify(DFBETA[j,])
+ }
> for (j in 1:3) boxplot(DFBETA[j,],style.bxp="old")
```

In generating the DFBETA statistics I noted that from equation (2)

$$\underline{\hat{\beta}} - \underline{\hat{\beta}}_{(-i)} = \frac{(X^TX)^{-1}\underline{x}_i E_i}{1-h_{ii}} = (X^TX)^{-1}\underline{x}_i \cdot \frac{E_i}{1-h_{ii}}$$

and if we string these together for all i, then the entire collection $(k+1)\times n$ DFBETA's

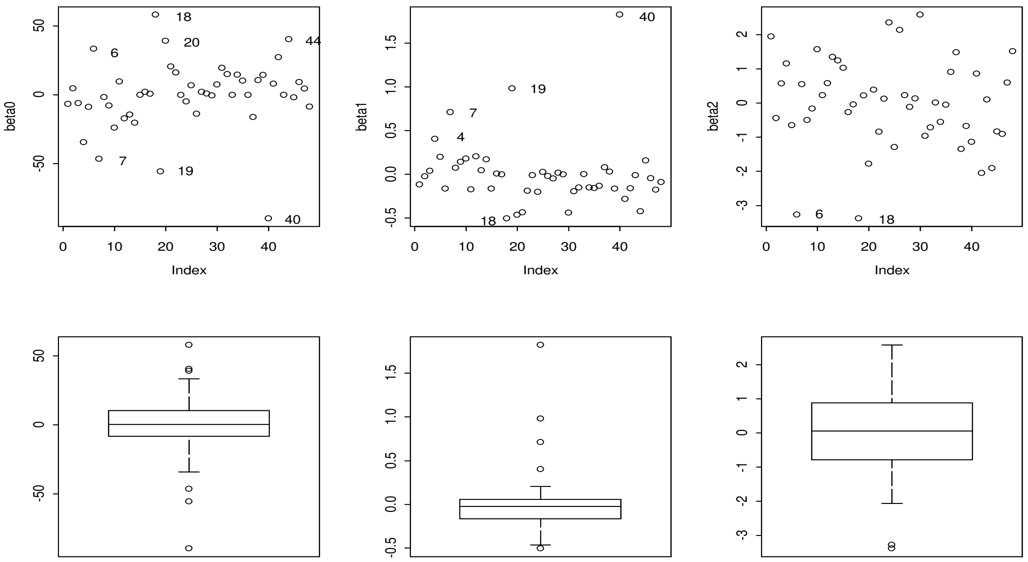


Figure 3: DFBETA ($\hat{\beta}_j - \hat{\beta}_{j(-i)}$) plots for `lm(formula = Fuel ~ Dlic + Tax, data = fuel48)`.

is given by

$$DFBETA = (X^T X)^{-1} X^T \begin{bmatrix} E_1/(1-h_{11}) & 0 & \cdots & 0 \\ 0 & E_2/(1-h_{22}) & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & E_n/(1-h_{nn}) \end{bmatrix}$$

The result is in Figure 3. We can identify the states that generated the largest changes in the beta coefficients (as determined by the boxplots in Figure 3):

```
> # For Beta0:
> dimnames(fuel48)[[1]][c(18,44,20,6,7,19,40)]
[1] "ND" "UT" "NE" "CN" "NY" "SD" "WY"
> # For Beta1:
> dimnames(fuel48)[[1]][c(40,19,7,4,18)]
[1] "WY" "SD" "NY" "MA" "ND"
> # For Beta2:
> dimnames(fuel48)[[1]][c(6,18)]
[1] "CN" "ND"
```

Observations 18, 19 and 40 (ND, SD and WY), which were identified in the Cook's plots, are also identified here as having the greatest influence on the coefficients. Of these, observation 40 (WY) clearly is the most extreme; observations 18 (ND) and 19 (SD) also have large influence but are not well-separated from the other observations in the index plots. □

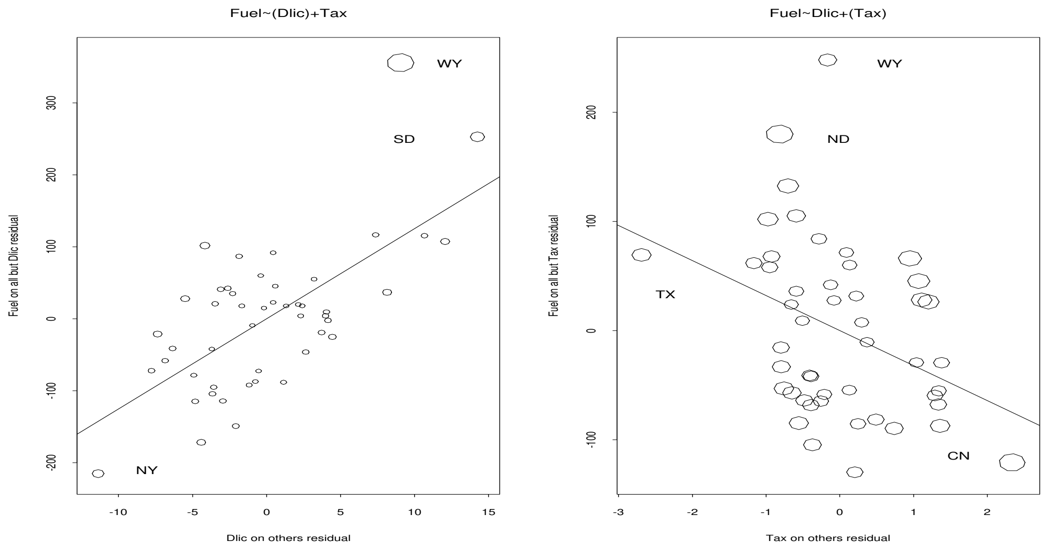
5 Partial Regression Plots

Partial regression plots, which we have used before to illustrate the effect of fitting a new regressor in a multiple regression equation, can also be useful for detecting influence, because they reduce the space to a simple two-dimensional space (regression of the residuals of Y on the other regressors onto the residuals of X_{new} onto the other regressors) in which outlyingness, leverage and influence of groups of points may sometimes be seen visually as in Figure 1.

```
> source("partreg.q")  # partial regression plot with dots
>                      # proportional to DFBETAS for the new beta
>                      # ...find it in http://.../~brian/707/data
> par(mfrow=c(1,2))

> partreg(xlst=c("Dlic","Tax"), yname="Fuel",
+   outname="Dlic", df=fuel48, prop=T, id=T, tcex=0.7, cex=0.8)

> partreg(xlst=c("Dlic","Tax"), yname="Fuel",
+   outname="Tax", df=fuel48, prop=T, id=T, tcex=0.7, cex=0.8)
```



6 Numerical Cutoffs vs. Graphical Displays

Although I have listed some numerical cutoffs for the various diagnostics above, these cutoffs are not to be trusted, for several reasons:

- Absolute cutoffs tend to not identify unusual data points as the sample size for the regression increases (the opposite of the usual complaint about classical hypothesis testing!): there is just too much "space" in n-dimensional space for any single point to pull the regression very far.

- Distributrion theory for the diagnostics is not always easy to obtain (as in Cook's D_i); nor may distribution theory even be available (as with h_{ii}, which are not realizations of random variables); and so cutoffs tend to be ad-hoc at best.
- Moreover even when distribution theory is available (as with the t-distribution of the studentized residuals E_i^*) the diagnostics are often statistically dependent, complicating any multiple-comparisons analysis.

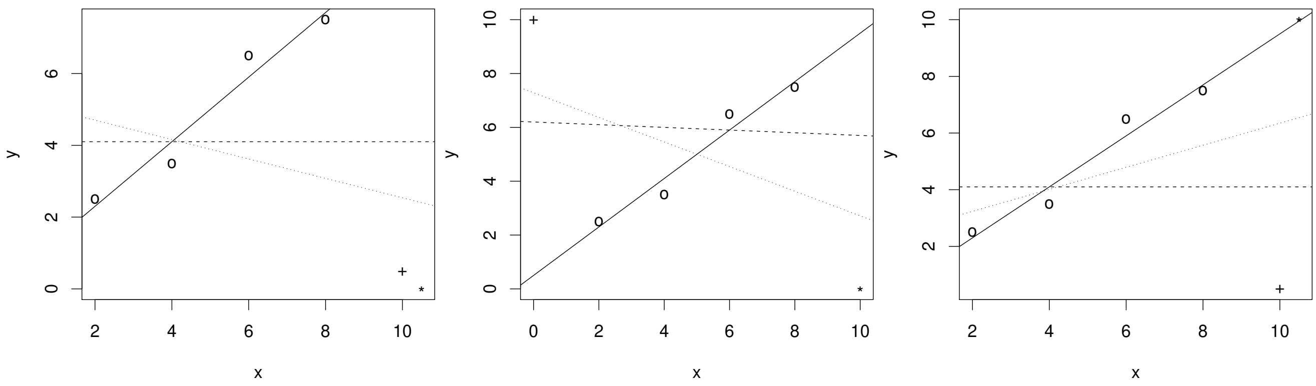
A generally better approach is to use a stem and leaf plot, time series plot, or other graphical device to inspect the most extreme 5% or so of the values of the diagnostic statistics, and then look at the data points identified thereby to make a qualitative judgement about what (if anything) to do about them.

7 Joint Influence

Several data points in a regression can interact in their influence on a regression fit. The basic lessons are:

- Sometimes removing one influential data point has little or no effect on the regression, because there is another influential point nearby.

- Sometimes removing a data point with high influence worsens the fit because the influence of that data point was cancelling out the influence of another data point, and the "hidden" influence of the second data point now degrades the regression.



Joint influence. The points marked "+" and "$*$" are both influential. The solid line is the regression without either influential point; the dashed line is the regression line with only the "+" point; and the dotted line is the regression line with both "+" and "$*$" points.

8 Diagnostics for Non-Linear Models

Beyond linear regression, there are many models that do not have the nice linearity properties

$$\left.\begin{array}{rcl} E[y|x] & = & x^T \underline{\beta} \\ \underline{\hat{\beta}} & = & (X^T X)^{-1} X^T \underline{y} \\ \underline{\hat{y}} & = & X(X^T X)^{-1} X^T \underline{y} \end{array}\right\} \tag{8}$$

and the simple relationship (1) that we exploited in developing leave-one-out statistics. For example,

- Log-linear models for contingency tables;
- Logistic Regression Models;
- Poisson Regression models;
- Survival Analysis models;
- Generalized Linear Models;
- and more general nonlinear models

are commonly used and do not enjoy these properties. However, most of these models involve a "regression surface"

$$E[y|\underline{x}] = g_{\underline{\beta}}(\underline{x})$$

that is "locally linear" (by Taylor's Theorem!); indeed the class of "generalized linear models" has the nice property

$$E[y|\underline{x}] = g(\underline{x}^T \underline{\beta})$$

Consequently

- Near the MLE $\hat{\beta}$ all of the properties in equations (8) hold approximately, with

$$X \approx \left. \frac{\partial E[\underline{y}]}{\partial \underline{\beta}} \right|_{\underline{\beta} = \underline{\hat{\beta}}}$$

- Consequently, approximate values of DFBETA's and Cook's D's can be calculated as usual;
- If your computer is "fast" and the data set is "not too large", it is even worthwhile computing all $n + 1$ models (the model with all the data, and the model leaving each data point out, in turn) and computing $DFBETA$'s and Cook's D's directly, using the definitions in equations (5) and (6) on page 8.