Case Deletion Diagnostics

- How do the estimated β's and/or estimated predictions $\hat{y}_j$ change when the i'th case is deleted?
- Do some cases strongly affect the fit?

Summary

- the **deleted residual**, $e_{(i)}$, is the residual using the prediction of $E[y_i]$ without case i
- $e_{(i)} = \frac{e_i}{1-h_{ii}}$
  - when the leverage is high, the deleted residual will be inflated
  - when the leverage is small, the deleted residual is close to the original residual.
- *variance of the deleted residual*: $\sigma^2/(1-h_{ii})$
- *prediction error sum of squares*: $PRESS = \sum_{i=1}^{n} e_{(i)}^2$
- *R^2 for prediction*: $R^2_{prediction} = 1 - \frac{PRESS}{SST}$
- *deleted residual sum of squares*: $SSE_{(i)} = SSE - \frac{e_i^2}{1-h_{ii}}$
- *deleted variance estimate*: $s^2_{(i)} = \frac{SSE_{(i)}}{n-k-2}$.

Studentized residuals

- the *externally studentized* residual is the deleted residual standardized with the deleted standard error

$$t_i = \frac{e_i/(1-h_{ii})}{s_{(i)}/\sqrt{1-h_{ii}}} = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}.$$

- Cases with standardized residual greater than 2 in absolute value should be examined by the data analyst.
- Externally studentized residuals are preferred to the basic residuals $y_i - \hat{y}_i$, as the studentized residuals are typically best at revealing cases which strongly influenced the fit.

Case Deletion Diagnostics

- the *deleted estimate of* $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i \frac{e_i}{1-h_{ii}}.$$

- *Cook's distance*:

$$D_i = \frac{1}{(k+1)s^2}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}) = \frac{(\hat{\mathbf{y}}_{(i)}-\hat{\mathbf{y}})'(\hat{\mathbf{y}}_{(i)}-\hat{\mathbf{y}})}{(k+1)s^2}$$

  - measures the change in the estimate of $\boldsymbol{\beta}$ when the ith case is deleted
  - and where $\hat{\mathbf{y}}_{(i)}$ are the predicted values based on all of the observations but the i'th, also measures the change in predicted values
  - the usual diagnostic is to flag an observation for which $D_i > F(.5, p, n-p) \approx 1$

Example: the wood beam data

- The the R function *influence.measures(lmoutput)* returns leverage values, Cook's distance, DFBETA's and DFFIT's.
- externally and internally studentized residuals are obtained for the linear model using *rstudent(woodlm.out)* and *rstandard(woodlm.out)*.
- Also shown are the leverage values,
- and deleted estimates of σ, obtained using *lm.influence(woodlm.out)$sigma*.

```
> cbind(rstudent(woodlm.out),rstandard(woodlm.out),
hat(model.matrix(woodlm.out)),lm.influence(woodlm.out)
$sigma)
Case  external     internal  leverage      s_(i)
1  -3.25407928 -2.11381331 0.4178935 0.1789230
2   0.26672584  0.28640388 0.2418666 0.2957614
3   0.18235097  0.19641804 0.4172806 0.2966886
4  -0.96116898 -0.96644039 0.6043904 0.2769510
5  -1.05851279 -1.04952172 0.2521824 0.2731008
6   2.20302617  1.76924218 0.1478688 0.2212052
7   0.50868523  0.53796494 0.2616385 0.2912946
8   0.43572812  0.46336607 0.1540321 0.2929114
9   0.26975149  0.28961404 0.3155106 0.2957218
10 -0.03324377 -0.03590407 0.1873364 0.2974822
```

- Note that cases 1 and 6 have large externally studentized residuals.
- These are also the cases with the smallest deleted estimates of σ.

- Cook's distance values are as follows.

```
> cooks.distance(woodlm.out)
           1            2            3            4
1.069240e+00 8.723020e-03 9.208950e-03 4.756415e-01
5            6
1.238171e-01 1.810604e-01
           7            8            9           10
3.418372e-02 1.303120e-02 1.288740e-02 9.905523e-05
```

- Cases 1, 4 and 6 give the largest values, and case 1 is above the threshhold.

The command *lm.influence(woodlm.out)$coefficients* returns the elements of $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$

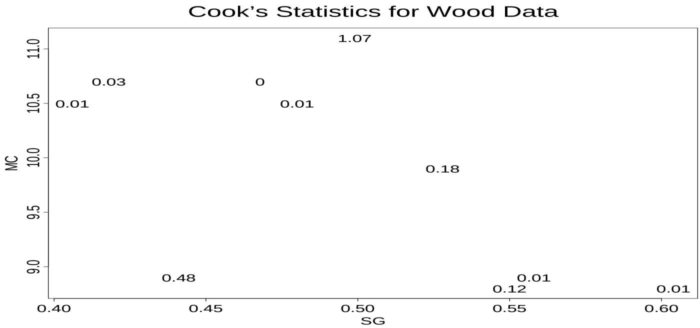
```
> lm.influence(woodlm)$coefficients:

  (Intercept)       SG       MC
1       2.7098 -1.7723 -0.1932
2       0.0458  0.0822 -0.0078
3      -0.0686  0.1887 -0.0018
4      -2.1091  1.6955  0.1242
5      -0.3672 -0.1320  0.0404
6      -0.6423  0.7481  0.0329
7       0.1203 -0.3082  0.0050
8      -0.1529  0.0642  0.0137
9       0.1669 -0.2550 -0.0031
10      0.0129 -0.0030 -0.0013
```

- To get the deleted parameter estimates

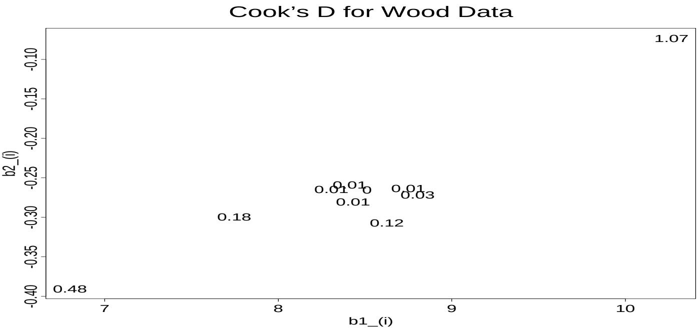
```
> wood.lm$coef-influence(wood.lm)$coefficients
   (Intercept)          SG          MC
1    7.5917355 10.2670553 -0.07314195
2    8.4489312 -0.3485564 10.30936123
3   -0.1976911 10.1128150  8.49650349
4   12.4106577  6.7992142 -0.39054922
5    8.8619301 -0.1343679 10.26112696
6    0.3759702  9.5534004  8.46184770
7   10.1812517  8.8028925 -0.27133686
8    8.6475688 -0.3305508 10.28786646
9   -0.4332042 10.5565539  8.49784592
10  10.2886005  8.4976812 -0.26505116

> lm(formula = Strength ~ SG + MC,data=data[-1,])
Coefficients:
(Intercept)          SG          MC
    7.59174     10.26706     -0.07314
```

- In case 1, the intercept is reduced, the slope for SG is increased, and the coefficient for MC is greatly reduced.
- In case 4, the intercept is increased, the slope for SG is decreased, and the slope for MC is greatly increased
- For case 6, the slope for SG is reduced.

- The following plot shows Cook's statistic in the space of predictors.
- The largest Cook's statistic is not the furthest from the centroid.



- The following plot shows the deleted parameter estimates $\hat{\beta}_{(i)}$ with the value of Cook's distance superimposed.



- We previously discussed adding a squared term in *moisture content*.
- The externally and internally studentized residuals, leverage values, deleted estimates of s and Cook's statistic are shown below.
- Note that the largest residual is for case 7, which also gives the smallest deleted estimate of σ.
- The largest value of Cook's distance is case 1, but its value is much less than 1.

```
> cbind(rstudent(woodlm2.out),rstandard(woodlm2.out),
hat(model.matrix(woodlm2.out)),lm.influence(woodlm2.
out)$sigma,cooks.distance(woodlm2.out))
```

Case	external	internal	leverage	sigma	Cook's
1	-0.8145812	-0.8384283	0.7657191	0.11170541	0.574386783
2	0.7064363	0.7379122	0.2418690	0.11336379	0.043429549
3	1.2342390	1.1836947	0.4241376	0.10408379	0.257992667
4	-0.3953950	-0.4265168	0.6469168	0.11707054	0.083326597
5	-1.5773574	-1.4119560	0.2836093	0.09714797	0.197311706
6	-0.1984146	-0.2165016	0.6163116	0.11842141	0.018822789
7	2.6278156	1.8655129	0.2662545	0.07704528	0.315709744
8	-0.7053329	-0.7368641	0.2304277	0.11337985	0.040644337
9	-0.3525067	-0.3814410	0.3371019	0.11743638	0.018497333
10	-0.1965493	-0.2144819	0.1876526	0.11843007	0.002656649

Derivations

- A trick is to delete the ith case by adding its indicator, $\mathbf{u}_i$ as a new predictor in the model!
- Let $u_{ij} = 1$ for $j = i$ and $u_{ij} = 0$ for $j \neq i$.
- Consider the expanded model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}_i\gamma + \boldsymbol{\epsilon}.$$

- This model gives a separate mean for the ith case

$$E[y_i] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \gamma = \mathbf{x}_i^T\boldsymbol{\beta} + \gamma.$$

- The sum of squares function is

$$SSE(\boldsymbol{\beta}, \gamma) = \sum_{j \neq i}^{n} (y_j - \mathbf{x}_j^T\boldsymbol{\beta})^2 + (y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \gamma)^2.$$

- The deleted estimate for $\boldsymbol{\beta}$ minimizes the first term, and is based on all cases but the ith.
- Denote these case deleted estimates $\hat{\boldsymbol{\beta}}_{(i)}$.
- The estimate for γ makes the second term zero, and is

$$\hat{\gamma} = y_i - \mathbf{x_i}^T \hat{\boldsymbol{\beta}}_{(i)}$$

the *deleted residual*, $e_{(i)}$, formed using the prediction of $E[y_i]$ without case i.

- Using the theory developed for adding a column to the $\mathbf{X}$ matrix, the deleted residual is

$$\begin{aligned}\hat{\gamma} &= [\mathbf{u}_i^T(\mathbf{I}-\mathbf{H})\mathbf{u}_i]^{-1}\mathbf{u}_i^T(\mathbf{I}-\mathbf{H})\mathbf{y} \\ &= \frac{e_i}{1-h_{ii}} = e_{(i)}.\end{aligned}$$

- That is, the deleted residual is just the original residual divided by one minus the leverage value
  - when the leverage is high, the deleted residual will be inflated
  - when the leverage is small, the deleted residual is close to the original residual.

- Deleted residuals are also called PRESS residuals, and are used to compute the prediction error sum of squares

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2$$

and the R^2 for prediction

$$R^2_{prediction} = 1 - \frac{PRESS}{SST}.$$

- Using our theory from before on adding variables to a model, we can determine that the deleted estimates of $\boldsymbol{\beta}$ are

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(i)} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{u}_i \frac{e_i}{1-h_{ii}}) \\ &= \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i \frac{e_i}{1-h_{ii}}.\end{aligned}$$

- The extra sum of squares for regression explained by the indicator is

$$
\begin{aligned}
SSR(\gamma|\boldsymbol{\beta}) &= \mathbf{y}^T(\mathbf{I}-\mathbf{H})u_i[u_i^T(\mathbf{I}-\mathbf{H})u_i]^{-1}u_i^T(\mathbf{I}-\mathbf{H})\mathbf{y} \\
&= \frac{e_i^2}{1-h_{ii}}.
\end{aligned}
$$

The increase in SSR is offset by a decrease in SSE so the deleted residual sum of squares is

$$SSE_{(i)} = SSE - \frac{e_i^2}{1-h_{ii}}.$$

Dividing this by $(n-1)-(k+1) = n-k-2$ gives the deleted variance estimate $s_{(i)}^2$.

The variance of the deleted residual is $\sigma^2/(1-h_{ii})$.

- Standardizing the deleted residual using the deleted standard error gives the *externally studentized* residual

$$t_i = \frac{e_i/(1-h_{ii})}{s_{(i)}/\sqrt{1-h_{ii}}} = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}.$$

- This differs from the *internally studentized residual*

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

  only through the estimates of standard error.

- Any case with standardized residual greater than 2 should be examined.

- The externally studentized residuals may reveal a case which has strongly influenced the fit.

Case Deletion Diagnostics

- There are numerous *case deletion diagnostics* which attempt to determine whether a case has strongly influenced the results.
- The most commonly used is *Cook's distance*

$$\begin{aligned} D_i &= \frac{1}{(k+1)s^2}(\hat{\beta}-\hat{\beta}_{(i)})^T\mathbf{X}^T\mathbf{X}(\hat{\beta}-\hat{\beta}_{(i)}) \\ &= \frac{1}{(k+1)s^2}(\hat{\mathbf{y}}-\hat{\mathbf{y}}_{(i)})^T(\hat{\mathbf{y}}-\hat{\mathbf{y}}_{(i)}) \\ &= \frac{e_i^2}{(k+1)s^2}\frac{h_{ii}}{(1-h_{ii})^2} \\ &= \frac{e_{(i)}^2}{(k+1)s^2} \end{aligned}$$

- D_i measures the change in the estimates of β and in the estimate of $E(\mathbf{y}) = \mathbf{X}\beta$ when the *i*th case is deleted.

- It also the scaled product of the squared deleted residual and the leverage value.
- It will be large when the ith case has a large deleted residual and also large leverage.
- It is often compared to $F(.5, p, n-p)$, which is close to 1. (Recall $p = k + 1$.)