Notes for 751-752
Sections 15-16

Martin Lindquist*

November 1, 2017

*Thanks to Brian Caffo and Ingo Ruczinski for supplying their notes, upon which much of this document is based.

14 Simultaneous Inference

Often we are interested in constructing a collection, or family, of confidence intervals each with a specific confidence level. We need to determine our level of confidence that all intervals simultaneously contain the true parameter value. Similarly, in many situations we have seek to perform multiple tests at the same time, rather than a single joint test. Here we focus on confidence intervals, but the methods described carry over to hypothesis testing as well.

Simultaneously making a large number of comparisons compounds the statistical uncertainty and introduces the need to adjust the individual confidence levels for multiple comparisons. In the context of confidence intervals, we need to distinguish between an individual confidence level and family confidence level. The individual confidence level is the confidence we have that any particular confidence interval contains the true parameter value. The familywise confidence level is the confidence we have that all the confidence intervals in a family of intervals simultaneously contain the true parameter value.

14.1 Multiple Comparisons: The Bonferroni Method

The Bonferroni procedure controls for multiple comparisons by adjusting the width of the margin of error. This makes the confidence interval wider for each individual contrast, but keeps the overall family confidence level at the desired level.

To illustrate, let us assume we are creating k confidence intervals for $\beta_1, \beta_2, \ldots \beta_k$. Suppose the j^{th} confidence interval has coverage probability $1-\alpha$ and we want the familywise confidence level to be $1-\alpha$. Let E_j be the event that the j^{th} includes β_j, and E_j^c be the complement. Then by definition,

$$\begin{aligned} 1-\alpha_f &= P(E_1 \cap E_2 \cap \ldots \cap E_k) \\ &= 1 - P(E_1^c \cup E_2^c \cup \ldots \cup E_k^c) \\ &\geq 1 - \sum_{j=1}^{k} P(E_j^c) \\ &= 1 - k\alpha. \end{aligned}$$

Thus, if the individual confidence level for two intervals is $1-\alpha = 0.95$, then we have a family confidence level of at least $1-2\alpha = 0.90$. In order to guarantee a family confidence level of at least 0.95 we instead need the individual confidence level to be $(1-0.05/2) = 0.975$.

We can ensure appropriate control if we set $\alpha = \alpha_f/k$. Using this approach, Bonferroni confidence intervals for $\beta_1, \beta_2, \ldots \beta_k$ are given by

$$\hat{\beta}_j \pm t_{n-p,1-\alpha_f/2k} s\sqrt{g_{jj}} \text{ for } j = 1, 2, \ldots k$$

where g_{jj} is the j^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

The main features of the Bonferroni method are that it is simple to use, and that it is conservative. The latter implies that the actual coverage probability is greater than claimed and the confidence intervals are wider than required.

Example: 14.1 One-way ANOVA $Y_{ij} = \mu + \tau_i + \varepsilon_{ij},\ (i = 1, \ldots, k).$

If we want confidence intervals for all pairwise comparisons $\{\tau_i - \tau_j, i \neq j\},$ there are $n_k = k \times (k-1)/2$ such comparisons. The Bonferroni method uses $\alpha_j = \alpha/n_k$. Hence with $\alpha = .05$ we have:

k	n_k	α_j
2	1	0.0500
3	3	0.0167
4	6	0.0083
5	10	0.0050
6	15	0.0033

14.2 Multiple Comparisons: Scheffé Method

Scheffé's method is based on the following theorem.

Theorem: 14.2 If $\mathbf{L}$ is positive definite, then

$$\max_{\mathbf{h} \neq \mathbf{0}} \left(\frac{(\mathbf{h}'\mathbf{b})^2}{\mathbf{h}'\mathbf{L}\mathbf{h}} \right) = \mathbf{b}'\mathbf{L}^{-1}\mathbf{b}.$$

Recall from our discussion of confidence ellipsoids that we can write:

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{ps^2} \sim F_{p,n-p}.$$

Applying Theorem 14.2 with $\mathbf{b} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and $\mathbf{L} = (\mathbf{X}'\mathbf{X})^{-1}$, we find that

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \max_{\mathbf{h} \neq \mathbf{0}} \left(\frac{(\mathbf{h}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^2}{\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}} \right).$$

Therefore,

$$P\left(\frac{1}{ps^2}\max_{\mathbf{h}\neq\mathbf{0}}\left(\frac{(\mathbf{h}'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))^2}{\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}}\right)\leq F_{p,n-p,1-\alpha}\right)=1-\alpha.$$

Equivalently,

$$P\left(\frac{(|\mathbf{h}'\hat{\boldsymbol{\beta}}-\mathbf{h}'\boldsymbol{\beta}|}{\sqrt{\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}}}\leq\sqrt{ps^2F_{p,n-p,1-\alpha}}\ \ \forall\,\mathbf{h}\in\mathbb{R}^p\right)=1-\alpha.$$

As a special case, we can write:

$$P\left(\frac{|\hat{\beta}_j-\beta_j|}{\sqrt{g_{jj}}}\leq\sqrt{ps^2F_{p,n-p,1-\alpha}}\ \ \forall\,1\leq j\leq p\right)\geq 1-\alpha.$$

Hence, simultaneous confidence intervals for $\beta_1,\beta_2,\ldots\beta_k$ are given by

$$\hat{\beta}_j\pm\sqrt{ps^2g_{jj}F_{p,n-p,1-\alpha}}\ \text{ for }\ j=1,2,\ldots k$$

More generally, we can express simultaneous confidence intervals for linear combinations of $\boldsymbol{\beta}$ as follows:

$$\mathbf{h}'\hat{\boldsymbol{\beta}}\pm\sqrt{pF_{p,n-p,1-\alpha}s^2\mathbf{h}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{h}}\quad\forall\mathbf{h}.$$

Example: 14.3 (Confidence bands for a regression surface).

Suppose we want simultaneous confidence intervals for the mean of the response variable y at a given set of predictor variables $\mathbf{x}'=(x_{i0},x_{i1},\ldots,x_{i,p-1})$, i.e. $E[y]=\mathbf{x}'\boldsymbol{\beta}$. Set $\mathbf{h}=\mathbf{x}$ and we obtain:

$$\mathbf{x}'\hat{\boldsymbol{\beta}}\pm\sqrt{pF_{p,n-p,1-\alpha}s^2\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}\quad\forall\mathbf{x}.$$

This gives us simultaneous confidence intervals for the mean of y at all values of the predictors. Plotted against the predictors, this yields a confidence band around the fitted model.

Example: 14.4 (Simple linear regression).

If $p=2$ we get the following confidence region for $\boldsymbol{\beta}=(\beta_0,\beta_1)'$:

$$\{\boldsymbol{\beta}:(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\leq 2F^{\alpha}_{2,n-2}s^2\}.$$

The simultaneous confidence intervals for β_0 and β_1 are given by

$$\begin{aligned}\hat{\beta}_0 &\pm \sqrt{\frac{2F_{2,n-2,1-\alpha}s^2\sum x_i^2/n}{\sum(x_i-\bar{x})^2}} \\ \hat{\beta}_1 &\pm \sqrt{\frac{2F_{2,n-2,1-\alpha}s^2}{\sum(x_i-\bar{x})^2}}\end{aligned}$$

where we have

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\sum(x_i-\bar{x})^2}\begin{pmatrix} \sum x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.$$

The confidence band for the regression line $\beta_0 + \beta_1 x$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm \sqrt{2F^{\alpha}_{2,n-2}s^2\frac{\sum(x_i-x)^2}{\sum(x_i-\bar{x})^2}}.$$

Verify that $\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} = \sum(x_i-x)^2/\sum(x_i-\bar{x})^2$. Note that the width of the confidence band depends on $\sum(x_i-x)^2/\sum(x_i-\bar{x})^2$, i.e. how far x is from $\bar{x}$.

15 Residuals and diagnostics

Now with some distributional results under our belt, we can discuss distributional properties of the residuals. Note that, as a non-full rank linear transformation of normals, the residuals are singular normal. When $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, the mean of the residuals is $\mathbf{0}$, variance of the residuals is given by

$$\text{Var}(\mathbf{e}) = \text{Var}\{(\mathbf{I} - \mathbf{H}_\mathbf{X})\mathbf{y}\} = \sigma^2(\mathbf{I} - \mathbf{H}_\mathbf{X}).$$

As a consequence, we see that the diagonal elements of $\mathbf{I} - \mathbf{H}_\mathbf{X} \geq 0$ and thus the diagonal elements of $\mathbf{H}_\mathbf{X}$ must be less than one. (A fact that we'll use later). In addition, we see that although the errors may have equal variance and be uncorrelated the residuals do not.

15.1 The hat matrix

Let us study the hat matrix

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

in more detail. Recall it is symmetric and idempotent. In the continuation we write the $(i,j)^{th}$ element of $\mathbf{H}$ as h_{ij}.

Since $\text{rank}(\mathbf{H}) = \text{trace}(\mathbf{H})$, it follows that $\sum h_{ii} = \tilde{p}$. If we assume there is an intercept in the model, then $\mathbf{H}\mathbf{J}_n = \mathbf{J}_n$, and hence $\sum_i h_{ij} = \sum_j h_{ij} = 1$.

Theorem: 15.1 Assume we have an intercept in the model. Let $\mathcal{X}$ be the $n \times (p-1)$ mean-centered design matrix (without the intercept term). Let $(\mathcal{X}'\mathcal{X})_{jk} = \sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$, and redefine $\mathbf{x}_i'$ to be the ith row of $\mathbf{X}$ without the one for the intercept. Then the following holds:

(a) $h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathcal{X}'\mathcal{X})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$. This means in particular that each h_{ii} is bounded below by $\frac{1}{n}$.

(b) Let r_i be the number of rows of $\mathbf{X}$ that are identical to its ith row. Then $h_{ii} \leq \frac{1}{r_i}$.

Note, as n increases h_{ii} tends to decrease. The term $(\mathbf{x}_i - \bar{\mathbf{x}})'(\mathcal{X}'\mathcal{X})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$ corresponds to a corresponds to a Mahalanobis distance, roughly estimating the distance between $\mathbf{x}_i$ and $\bar{\mathbf{x}}$.

Example: 15.2 Consider the case of simple linear regression. Here $\mathcal{X} = (x_1, x_2, \ldots x_n)'$. Then,

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_k (x_k - \bar{x})^2}$$

Note $h_{ii} = \frac{1}{n}$ if and only if $x_i = \bar{x}$.

15.2 Studentized Residuals

It is important to note that the variance of the residuals is not constant, as

$$Var(e_i) = \sigma^2(1 - h_{ii}).$$

Since $h_{ii} \leq 1$, the variance will be small if h_{ii} is close to 1. In general, this will be true if the observation lies far away from the mean. This can be problematic, as the model is less likely to hold for these observations.

Another problem with the residuals is that they have the units of $\mathbf{y}$ and thus are not comparable across experiments. Taking

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

i.e., standardizing the residuals by their estimated standard deviation, solves these problems. However, the resulting quantities are not comparable to t-statistics since the numerator elements (the residuals) are not independent of s^2. The residuals standardized in this way are called "studentized" residuals. Studentized residuals are a standard part of most statistical software.

15.3 Coding example

```
> data(mtcars)
> y = mtcars$mpg
> x = cbind(1, mtcars$hp, mtcars$wt)
> n = nrow(x); p = ncol(x)
> hatmat =  x %*% solve(t(x) %*% x) %*% t(x)
> residmat =  diag(rep(1, n)) - hatmat
> e = residmat %*% y
> s = sqrt(sum(e^2) / (n - p))
> rstd = e / s / sqrt(diag(residmat))
> # compare with rstandard, r's function
> # for calculating standarized residuals
> cbind(rstd, rstandard(lm(y ~ x - 1)))
          [,1]        [,2]
1  -1.01458647 -1.01458647
2  -0.62332752 -0.62332752
3  -0.98475880 -0.98475880
4   0.05332850  0.05332850
5   0.14644776  0.14644776
6  -0.94769800 -0.94769800
...
```

15.4 PRESS residuals

Consider the model $\mathbf{y} \sim N(\mathbf{W}\boldsymbol{\gamma}, \sigma^2\mathbf{I})$ where $\gamma = [\boldsymbol{\beta}' \ \Delta_i]$, $\mathbf{W} = [\mathbf{X} \ \boldsymbol{\delta}_i]$ where $\boldsymbol{\delta}_i$ is a vector of all zeros except a 1 for row i. This model has a shift in position i, for example if there is an outlier at that position. This is called the mean-shift outlier model. The least squares criterion can be written as

$$\sum_{k \neq i} \left(y_k - \sum_{j=1}^{p} x_{kj}\beta_j \right)^2 + \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \Delta_i \right)^2 . \tag{1}$$

Consider holding $\boldsymbol{\beta}$ fixed, then we get that the estimate of Δ_i must satisfy

$$\Delta_i = y_i - \sum_{j=1}^{p} x_{ij}\beta_j$$

and thus the right hand term of (1) is 0. Then we obtain $\boldsymbol{\beta}$ by minimizing

$$\sum_{k \neq i} \left(y_k - \sum_{j=1}^{p} x_{kj}\beta_j \right)^2 .$$

Therefore $\hat{\boldsymbol{\beta}}$ is exactly the least squares estimate having deleted the i^{th} data point; notationally, $\hat{\boldsymbol{\beta}}_{(i)}$. Thus, $\hat{\delta}_i$ is a form of residual obtained when deleting the i^{th} point from the fitting then comparing it to the fitted value,

$$\hat{\Delta}_i = y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_{(i),j}.$$

Notice that the fitted value at the i^{th} data point is then $\sum_{j=1}^{p} x_{ij}\hat{\beta}_{(i),j} + \hat{\Delta}_i = \ y_i$ and thus the residual is zero. The term $\hat{\Delta}_i$ is called the PRESS residual, the difference between the observed value and the fitted value with that point deleted.

Since the residual at the i^{th} data point is zero, the estimated variance from this model is exactly equal to the variance estimate having removed the i^{th} data point. The t test for δ_i is then a form of standardized residual, that exactly follows a t-distribution under the null hypothesis that $\delta_i = 0$.

15.5 Computing PRESS residuals

It is interesting to note that PRESS residuals don't actually require recalculating the model with the i^{th} datapoint deleted. Let $\mathbf{X}' = [\mathbf{z}_1 \ \ldots \ \mathbf{z}_n]$ so that $\mathbf{z}_i$ is the i^{th} row of the matrix $\mathbf{z}$ (hence column i of $\mathbf{z}'$). We use $\mathbf{z}$ for the rows, since we've already reserved $\mathbf{x}$ for the columns of $\mathbf{X}$. Notice, then that

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i'.$$

Thus, $\mathbf{X}'_{(i)}\mathbf{X}_{(i)}$, the x transpose x matrix with the i^{th} data point deleted is simply

$$\mathbf{X}'_{(i)}\mathbf{X}_{(i)} = \mathbf{X}'\mathbf{X} - \mathbf{z}_i\mathbf{z}'_i.$$

We can appeal to the Sherman, Morrison, Woodbury theorem for the inverse

$$(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}_i\mathbf{z}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - \mathbf{z}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}_i}$$

Define h_{ii} as diagonal element i of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ which is equal to $\mathbf{z}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}_i$. (To see this, pre and post multiply this matrix by a vector of zeros with a one in the position i, an operation which grabs the i^{th} diagonal entry.) Furthermore, note that $\mathbf{X}'\mathbf{y} = \sum_{i=1}^{n}\mathbf{z}_i y_i$ so that

$$\mathbf{X}'_{(i)}\mathbf{y}_{(i)} = \mathbf{X}'\mathbf{y} - \mathbf{z}_i y_i.$$

Then we have that the predicted value for the i^{th} data point where it was not used in the fitting is:

$$\begin{aligned}
\hat{y}_{(i),i} &= \mathbf{z}'_i(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)} \\
&= \mathbf{z}'_i\left((\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}_i\mathbf{z}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}}\right)(\mathbf{X}'\mathbf{y} - \mathbf{z}_i y_i) \\
&= \hat{y}_i + \frac{h_{ii}}{1 - h_{ii}}\hat{y}_i - h_{ii}y_i - \frac{h_{ii}^2 y_i}{1 - h_{ii}} \\
&= \frac{\hat{y}_i}{1 - h_{ii}} + y_i - \frac{y_i}{1 - h_{ii}}
\end{aligned}$$

So that we wind up with the equality:

$$y_i - \hat{y}_{(i),i} = \frac{y_i - \hat{y}_i}{1 - h_{ii}} = \frac{e_i}{1 - h_{ii}}$$

In other words, the PRESS residuals are exactly the ordinary residuals divided by $1 - h_{ii}$.

15.6 Externally studentized residuals

It's often useful to have standardized residuals where a data point in question didn't influence the residual variance. The normalized PRESS residuals are, as seen in 15.4. However, the PRESS residuals are leave one out residuals, and thus the i^{th} point was deleted for the fitted value. An alternative strategy is to normalize the ordinary residuals by dividing by a standard deviation estimate calculated with the i^{th} data point deleted. That is,

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

In this statistic, observation i hasn't had the opportunity to impact the variance estimate. Note the internally and externally studentized residuals are monotonically related through

$$t_i = r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}}.$$

Given that the PRESS residuals are $\frac{e_i}{1-h_{ii}}$, their variance is $\sigma^2/\sqrt{1-h_{ii}}$. Then we have that the press residuals normalized (divided by their standard deviations) are

$$\frac{e_i}{\sigma\sqrt{1-h_{ii}}}$$

If we use the natural variance estimate for the PRESS residuals, the estimated variance calculated with the i^{th} data point deleted, then the normalized PRESS residuals are the same as the externally standardized residuals. As we know from above, these also arise out of the t-test for the mean shift outlier model from Section 15.4.

15.7 Coding example

First let's use the `swiss` dataset to show how to calculate the ordinary residuals and show that they are the same as those output by `resid`.

```
> y = swiss$Fertility
> x = cbind(1, as.matrix(swiss[,-1]))
> n = nrow(x); p = ncol(x)
> hatmat = x %*% solve(t(x) %*% x) %*% t(x)
> ## ordinary residuals
> e = (diag(rep(1, n)) - hatmat) %*% y
> fit = lm(y ~ x)
> ## show that they're equal by taking the max absolute difference
> max(abs(e - resid(fit)))
[1] 4.058975e-12
```

Next, we calculate the standardized residuals and show how to get them automatically with `rstandard`.

```
> ## standardized residuals
> s = sqrt(sum(e ^ 2) / (n - p))
> rstd = e / s / sqrt(1 - diag(hatmat))
> ## show that they're equal by taking the max absolute difference
> max(abs(rstd - rstandard(fit)))
[1] 6.638023e-13
```

Next, let's calculate the PRESS residuals both by leaving out the i^{th} observation (in this case observation 6) and by the shortcut formula.

```
> i = 6
> yi = y[i]
> yihat = predict(fit)[i]
> hii = diag(hatmat)[i]
> ## fitted model without the ith data point
> y.minus.i = y[-i]
> x.minus.i = x[-i,]
> beta.minus.i = solve(t(x.minus.i) %*% (x.minus.i)) %*% t(x.minus.i) %*% y.minus.i
> yhat.i.minus.i = sum(x[i,] * beta.minus.i)
> pressi = yi - yhat.i.minus.i
> c(pressi, e[i] / (1 - hii))
          Porrentruy
 -17.96269  -17.96269
```

Now show that the `rstudent` (externally studentized) residuals and normalized PRESS residuals are the same.

```
> ## variance estimate with i deleted
> e.minus.i = y.minus.i - x.minus.i %*% beta.minus.i
> s.minus.i = sqrt(sum(e.minus.i ^ 2) / (n - p  - 1))
> ## show that the studentized residual is the PRESS residual standardized
> ei / s.minus.i / sqrt(1 - hii)
Porrentruy
 -2.367218
> rstudent(fit)[i]
        6
-2.367218
```

Finally, show that the mean shift outlier model residuals give the PRESS and the rstudent residuals.

```
> delta = rep(0, n); delta[i] = 1
> w = cbind(x, delta)
> round(summary(lm(y ~ w - 1))$coef, 3)
                 Estimate Std. Error t value Pr(>|t|)
w                  65.456     10.170   6.436    0.000
wAgriculture       -0.210      0.069  -3.067    0.004
wExamination       -0.323      0.242  -1.332    0.190
```

```
wEducation          -0.895      0.174  -5.149    0.000
wCatholic           0.113       0.034   3.351    0.002
wInfant.Mortality   1.316       0.376   3.502    0.001
wdelta            -17.963       7.588  -2.367    0.023
```

So notice that the the estimate for `wdelta` is the PRESS residual while the `t value` is the externally studentized residual.

15.8 Leverage

The term h_{ii} from the hat matrix is called the leverage of the i^{th} observation. The leverage of a observation measures its ability to move the regression model all by itself by simply moving in the y-direction. The leverage measures the amount by which the predicted value would change if the observation was shifted one unit in the y-direction. The leverage is always between 0 and 1. A point with close to zero leverage has little effect on the regression model. If a point has leverage equal to 1 the line must follow the observation perfectly. The h_{ii} depends only on $\mathbf{X}$, though a knowledge of $\mathbf{y}$ is required for a full interpretation.

15.8.1 Cook's Distance

Cooks distance measures the aggregate influence of the i^{th} value on all n fitted values. It is defined by

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{\tilde{p}\hat{\sigma}^2} = \frac{(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}}_{(i)} - \hat{\mathbf{Y}})}{\tilde{p}\hat{\sigma}^2},$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ are the parameter estimates obtained after deleting observation i, and $\hat{\mathbf{Y}}_{(i)}$ are the corresponding fitted values.

An alternative expression for the Cook's distance is $D_i = \frac{r_i^2}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right)$. The value of D_i depends on two functions, the size of the residuals e_i and the leverage value h_{ii}. Hence, an observation can be influential by having a large residual and/or a large leverage. Typically, points with D_i greater than 1 are classified as influential.