Advanced Methods in Biostatistics II

Lecture 4

November 2, 2017

Linear model

- Consider the linear model

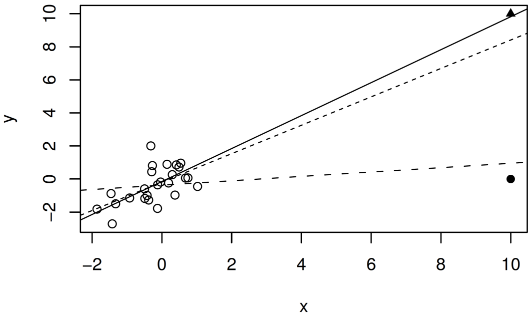
$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

  where $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$.

- Today we discuss various methods for testing the adequacy and validity of the model.

Notation

- An outlier is a point that does not fit the current model.
- An influential point is one whose removal from the dataset would cause a large change in the fit.



Residuals

- The error term ε is unobservable unless β is known.
- Therefore we estimate ε for a given sample using the residual vector:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

- The residuals contain information necessary for performing model diagnostics.

Properties of the Residuals

Properties

1. $E(\mathbf{e}) = \mathbf{0}$
2. $Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$
3. $Cov(\mathbf{e}, \mathbf{y}) = \sigma^2(\mathbf{I} - \mathbf{H})$
4. $Cov(\mathbf{e}, \hat{\mathbf{y}}) = \mathbf{0}$
5. $\sum_{i=1}^{n} e_i/n = 0$
6. $\mathbf{e}'\hat{\mathbf{y}} = 0$
7. $\mathbf{e}'\mathbf{X} = \mathbf{0}'$

Properties of the Residuals

- Note that the residual vector has the same mean as the error term ε, but a different variance-covariance matrix.
- In particular, note that the residuals are not independent.
- However, if n is large the h_{ij}'s tend to be small (for $i \neq j$), and the dependence won't have a significant effect on model diagnostics.

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.
- We write the $(i, j)^{th}$ element of $\mathbf{H}$ as h_{ij}.

Properties of the Hat Matrix

Theorem

$tr(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = p.$

Properties of the Hat Matrix

Theorem

Let $\mathbf{X}$ be a design matrix containing an intercept term. Let $\mathcal{X}$ be the $n \times (p-1)$ mean-centered design matrix (without the intercept). 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 i^{th} row of $\mathcal{X}$. Then the following holds:

(a) $h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathcal{X}'\mathcal{X})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$.

(b) $(1/n) \leq h_{ii} \leq 1$ for $i = 1, \ldots n$.

Properties of the Hat Matrix

- 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 Mahalanobis distance, providing an estimate of the distance between $\mathbf{x}_i$ and $\bar{\mathbf{x}}$.

Example

- Consider the case of simple linear regression.
- Here $\mathcal{X} = (x_1 - \bar{x}, x_2 - \bar{x}, \ldots x_n - \bar{x})'$.
- 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 $x_i = \bar{x}$.

Properties of the Residuals

- We are often interested in detecting outliers in the data.
- One approach is to plot $\mathbf{e}$ against $\hat{\mathbf{y}}$.
- However, it is important to note that the variance of the residuals is not constant, as

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

Properties of the Residuals

- 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 may be less likely to hold for these observations.

Studentized residuals

- To circumvent this issue, we often standardize the residuals as follows:

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

- These are called the internally “studentized” residuals.

Studentized residals

- Note the resulting quantities are not directly comparable to a t-statistic as the numerator elements (i.e., the residuals) are not independent of s^2.
- In contrast to ordinary residuals, studentized residuals have constant variance.
- Studentized residuals are a standard part of most statistical software.

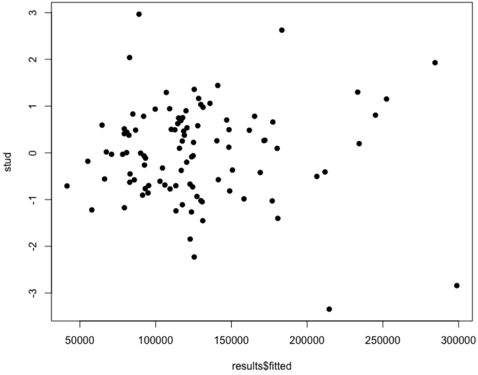
R code

```
> Housing = read.table("housing.txt", header=TRUE)
> Housing
    Taxes   Bedrooms        Baths  Price   Size   Lot
1     1360        3           2.0  145000  1240  18000
2     1050        1           1.0   68000   370  25000
3     1010        3           1.5  115000  1130  25000
..
99    1770        3           2.0   88400  1560  12000
100   1430        3           2.0  127200  1340  18000


> results = lm(Price ˜ Taxes + Size, data=Housing)
```

R code

```
> stud = rstandard(results)
> plot(results$fitted,stud, pch=19)
```



Deleted residuals

- Another approach is to compute the deleted residuals.
- This is the i^{th} residual obtained using a fitted model based on using all the data except the i^{th} observation.
- In the event that the i^{th} observation is influential, the fitted value will not be influenced by this observation and will tend to give a larger residual making it easier to detect.

Deleted residuals

- The deleted residual (or PRESS residual) for the i^{th} case is defined as

$$
\begin{aligned}
d_i &= y_i - \hat{y}_{(i)} \\
&= y_i - \mathbf{x}_i' \hat{\beta}_{(i)}
\end{aligned}
$$

- Here $\hat{y}_{(i)}$ denotes the fitted value, computed without the i^{th} observation, at $\mathbf{x}$ levels corresponding to that observation.

- Similarly, $\hat{\beta}_{(i)}$ denotes the estimated parameter, computed without the i^{th} observation.

Deleted residuals

- It is important to note that computing d_i doesn't actually require fitting the model with the i^{th} observation deleted.
- To illustrate, 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 $\mathbf{z}'$).
- Here we use $\mathbf{z}$ for the rows, since we've already reserved $\mathbf{x}$ for the columns of $\mathbf{X}$.

Deleted residuals

- Let us define $\mathbf{X}_{(i)}$ and $\mathbf{y}_{(i)}$ as the matrix $\mathbf{X}$ and the vector $\mathbf{y}$, respectively, with the i^{th} observation deleted.
- Note, that

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

- Thus,

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

Sherman-Morrison formula

Theorem

$$(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}}$$

Deleted residuals

- According to the Sherman-Morrison formula:

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

- Note $h_{ii} = \mathbf{z}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}_i$.
- Furthermore, note that $\mathbf{X}'\mathbf{y} = \sum_{i=1}^{n} \mathbf{z}_i y_i$ so that

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

Deleted residuals

- Now,

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

Deleted residuals

- Hence, it holds that

$$\hat{\beta} - \hat{\beta}_{(i)} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}_i}{1 - h_{ii}} e_i$$

- This will be useful for a later derivation.

Deleted residuals

- Using these results, we find that

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

- In other words, the deleted residuals are exactly the ordinary residuals divided by $1 - h_{ii}$.
- The deleted residuals are often used in model selection.

Externally studentized residuals

- An alternative approach for standardizing the residuals is given by:

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

- Here $s_{(i)}$ is the standard deviation estimated without using the i^{th} observation.
- These are called the externally “studentized” residuals.

Externally studentized residuals

- In this statistic, observation i doesn't impact the variance estimate.
- They follow a t_{n-p-1} distribution, which makes them useful for testing whether an observation is an outlier.
- Note the internally and externally studentized residuals are monotonically related through

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

Deleted vs. Externally studentized residuals

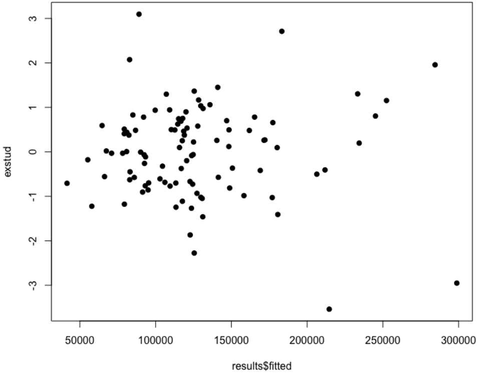
- Given that the deleted residuals are $\frac{e_i}{1-h_{ii}}$, their variance is given by $\sigma^2/1 - h_{ii}$.
- Thus, the normalized deleted residuals are

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

- If we use the estimated variance calculated with the i^{th} data point deleted, then the normalized deleted residuals are equal to the externally standardized residuals

R code

```
> exstud = rstudent(results)
> plot(results$fitted,exstud, pch=19)
```



Leverage

- The leverage of an observation measures the amount by which the predicted value would change if the observation is shifted one unit in the y-direction.
- The leverage is always between 0 and 1.
- A point with leverage close to zero has little effect on the regression model.
- If a point has leverage equal to 1 the line must follow the observation perfectly.

Leverage

- Recall that the fitted values can be written:

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

  or, alternatively

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n$$

  for $i = 1, 2, \ldots n$.

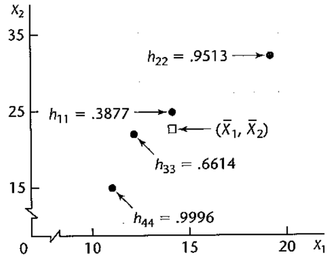
- Hence, the term h_{ii} is the leverage for the i^{th} observation.

Leverage

- If the i^{th} observation is an outlier in terms of its $\mathbf{x}$ observation it has a large leverage value.
- Since the leverage is a function only of $\mathbf{x}$ it measures the role of $\mathbf{x}$ in determining how y_i effects the fitted value.
- Outliers in the x-direction tend to have higher leverage values and thus a larger effect on the fitted regression function.

Illustration

FIGURE 10.6 Illustration of Leverage Values as Distance Measures—Table 10.2 Example.



Leverage

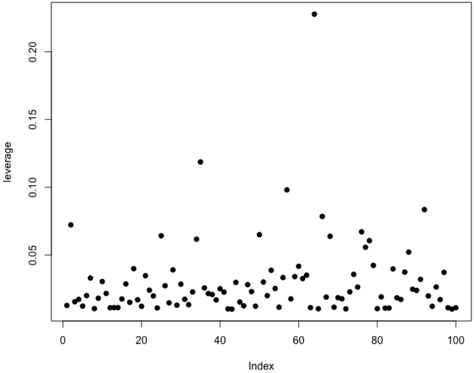
- Leverage is typically considered to be large if it more than twice as large as the mean leverage value,

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{p}{n}$$

- Another common guideline is that h_{ii} exceeding 0.5 indicates high leverage, while values between 0.2 and 0.5 indicate moderate leverage.

R code

```
> lev = hatvalues(results)
> plot(lev, pch=19, ylab = 'leverage')
```

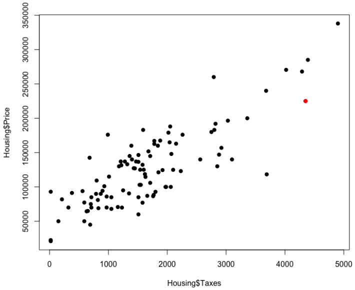


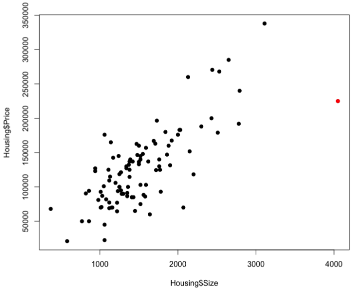
R code

```
> Housing[lev >0.2,]
     Taxes    Bedrooms   Baths   Price      Size     Lot
64    4350        3        3     225000     4050    35000
```

R code

```
> plot(Housing$Taxes,Housing$Price, pch =19)
> points(Housing[64,]$Taxes,Housing[64,]$Price, pch=19, col='red')
> plot(Housing$Size,Housing$Price, pch =19)
> points(Housing[64,]$Size,Housing[64,]$Price, pch=19, col='red')
```





Cook's Distance

- Cooks distance is a measure of the aggregate influence of the i^{th} observation on all n fitted values.
- It is defined as follows:

$$
\begin{aligned}
D_i &= \frac{(\hat{\beta} - \hat{\beta}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\beta} - \hat{\beta}_{(i)})}{ps^2} \\
&= \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{ps^2}.
\end{aligned}
$$

Cook's Distance

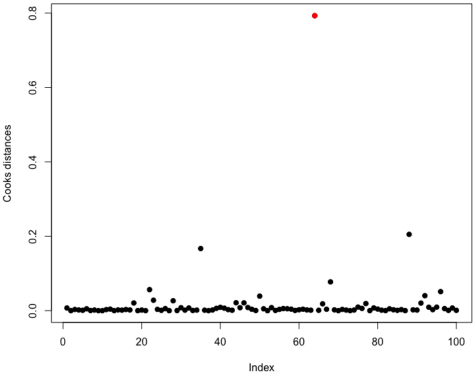
- An alternative expression is given by:

$$D_i = \frac{r_i^2}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right).$$

- 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.

R code

```
> cook = cooks.distance(results)
> plot(cook, pch=19, ylab="Cooks distances")
> points(64,cook[64], pch=19, col='red')
```



R code

```
> par(mfrow=c(2,2))
> plot(results)
```