Stat 206: Linear Models

Lecture 14

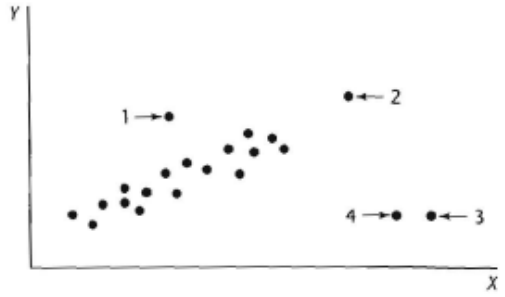
Nov. 18, 2015

Outlying Cases

Often, a data set contains some cases that are outlying or extreme, i.e., data points that are far away from the rest of the data.

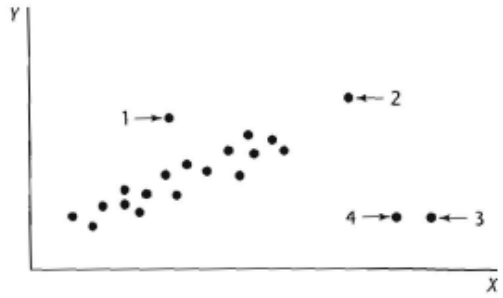
- A case may be outlying with respect to its Y value and/or its X value(s).
- Some (but not necessarily all) outlying cases may have an unduly strong influence on the fitted regression function – *influential cases*.
- An important step in regression analysis is to identify outlying cases and to investigate their effects in order to decide whether they should be retained or eliminated, and if retained, whether we should reduce their influence in the fitting process.

Examples of Outlying Cases



- Case 1 is outlying in , but influential: There are a few other cases with similar X values and they will keep the fitted regression line from being distorted too much by Case 1.
- Case 2 is outlying in , but influential: Its Y value is consistent with the regression relation suggested by other cases.
- Cases 3 and 4 are influential: They are outlying in X and their Y values are not consistent with the regression relation suggested by other cases.

Examples of Outlying Cases



- Case 1 is outlying in Y, but not very influential: There are a few other cases with similar X values and they will keep the fitted regression line from being distorted too much by Case 1.
- Case 2 is outlying in X, but also not very influential: Its Y value is consistent with the regression relation suggested by other cases.
- Cases 3 and 4 are likely to be very influential: They are outlying in X and their Y values are not consistent with the regression relation suggested by other cases.

Identify Outlying Cases

- With one or two X variables, outlying cases can be identified by simple graphical tools such as scatter plots, boxplots, etc.
- With multiple X variables, univariate outliers may not be extreme under the multivariate context, and, conversely, multivariate outliers may not be detectable using single- or bivariate- analyses. Often:
  - Outlying Y observations are identified through examining residuals.
  - Outlying X observations are identified through examining the diagonal elements h_{ii} of the hat matrix, called *leverage* values.

Residuals

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}, \quad \textit{assume}, \quad \textit{Var}(\mathbf{Y}) = \sigma^2\mathbf{I}_n.$$

- $\sigma^2\{\mathbf{e}\} =$, $\mathbf{s}^2\{\mathbf{e}\} =$.
  If the model is correct, then $\mathbf{E}\{\mathbf{e}\} =$.
- Variance of the *ith* residual:

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii}), \quad i = 1, \cdots, n.$$

  - Residual variances are in between
    .
  - The cases with larger h_{ii} have residual variances.
  - Residual variances goes to σ^2 as $n \to \infty$ (since $h_{ii} \to 0$).

Residuals

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}, \quad \textit{assume}, \quad \textit{Var}(\mathbf{Y}) = \sigma^2\mathbf{I}_n.$$

- $\sigma^2\{\mathbf{e}\} = \sigma^2 \times (\mathbf{I}_n - \mathbf{H})$, $\mathbf{s}^2\{\mathbf{e}\} = \textit{MSE} \times (\mathbf{I}_n - \mathbf{H})$. **If the model is correct**, then $\mathbf{E}\{\mathbf{e}\} = \mathbf{0}_n$.
- Variance of the *ith* residual:

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii}), \quad i = 1, \cdots, n.$$

  - Residual variances are in between 0 and σ^2.
  - The cases with larger h_{ii} have smaller residual variances.
  - Residual variances goes to σ^2 as $n \to \infty$ (since $h_{ii} \to 0$)

Studentized Residuals

(Internally) Studentized residuals:

$$r_i = \frac{e_i}{s\{e_i\}} = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}, \quad i = 1, \cdots, n.$$

- Studentized residuals have (roughly) constant variance across cases and thus are comparable to each other.
- In the R function *plot.lm()*, the residuals QQ (which=2), scale-location (which=3) and residuals vs. leverage (which=5) plots, studentized residuals are used.

Deleted Residuals

In order to make residuals more effective in detecting outlying Y observations, when calculating the residual of the *ith* case, we use the fitted regression function based on

.

- Such residuals are called *deleted residuals*:

$$d_i := Y_i - \widehat{Y}_{i(i)}, \quad i = 1, \cdots, n.$$

- If Y_i is far outlying, then the fitted regression function based on all cases may be by the *ith* case to be to Y_i such that e_i will be and fail to detect Y_i as outlying.
- If the *ith* case is excluded in fitting the regression function, then the fitted value for the *ith* case would be influenced by Y_i and the corresponding residual is more likely to detect Y_i if it is outlying.

Deleted Residuals

In order to make residuals more effective in detecting outlying Y observations, when calculating the residual of the *ith* case, we use the fitted regression function based on all but the *ith* case.

- Such residuals are called *deleted residuals*:

$$d_i := Y_i - \widehat{Y}_{i(i)}, \quad i = 1, \cdots, n.$$

- If Y_i is far outlying, then the fitted regression function based on all cases may be "dragged" by the *ith* case to be close to Y_i such that e_i will be small and fail to detect Y_i as outlying.
- If the *ith* case is excluded in fitting the regression function, then the fitted value for the *ith* case would not be influenced by Y_i and the corresponding residual is more likely to detect Y_i if it is outlying.

The deleted residual for the *ith* case equals to:

- The larger is h_{ii}, the the deleted residual d_i compared with the ordinary residual e_i.
- Sometimes deleted residuals will identify outlying Y observations not identified by ordinary residuals (when h_{ii} large) and sometimes they result in same identification as ordinary residuals (when h_{ii} small).

The deleted residual for the *ith* case equals to:

$$d_i = \frac{e_i}{1 - h_{ii}}, \quad i = 1, \cdots n.$$

- The larger is h_{ii}, the larger the deleted residual d_i compared with the ordinary residual e_i.
- Sometimes deleted residuals will identify outlying Y observations not identified by ordinary residuals (when h_{ii} large) and sometimes they result in same identification as ordinary residuals (when h_{ii} small).

Studentized Deleted Residuals

The *studentized deleted residuals (a.k.a. externally studentized residuals)* are defined as:

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{d_i}{\sqrt{MSE_{(i)}/(1 - h_{ii})}}, \quad i = 1, \cdots, n,$$

where $MSE_{(i)}$ is the MSE of the regression fit based on cases excluding case i.

- Studentized deleted residuals can be computed from the regression fit based on all cases:

$$t_i = e_i \sqrt{\frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2}}, \quad i = 1, \cdots, n.$$

Identify Outlying Y by Residual

H_0: The model is correct and all cases follow the model. Then under H_0:

$$t_i = \frac{d_i}{s\{d_i\}} \underset{H_0}{\sim} t_{(n-p-1)}, \quad i = 1, \cdots, n.$$

The d.f. is $n - p - 1$ since the deleted residuals are from regression fits based on $n - 1$ cases.

- Outlying Y observations are identified by large $|t_i|$.
- Since we are testing for n cases, we need to adjust for *multiple comparison*.
- Given significance level α , the **Bonferroni's procedure** controls the family-wise-typeI-error-rate at α by identifying cases with

  $$|t_i| > t(1 - \alpha/(2n); n - p - 1)$$

  as outlying Y observations.

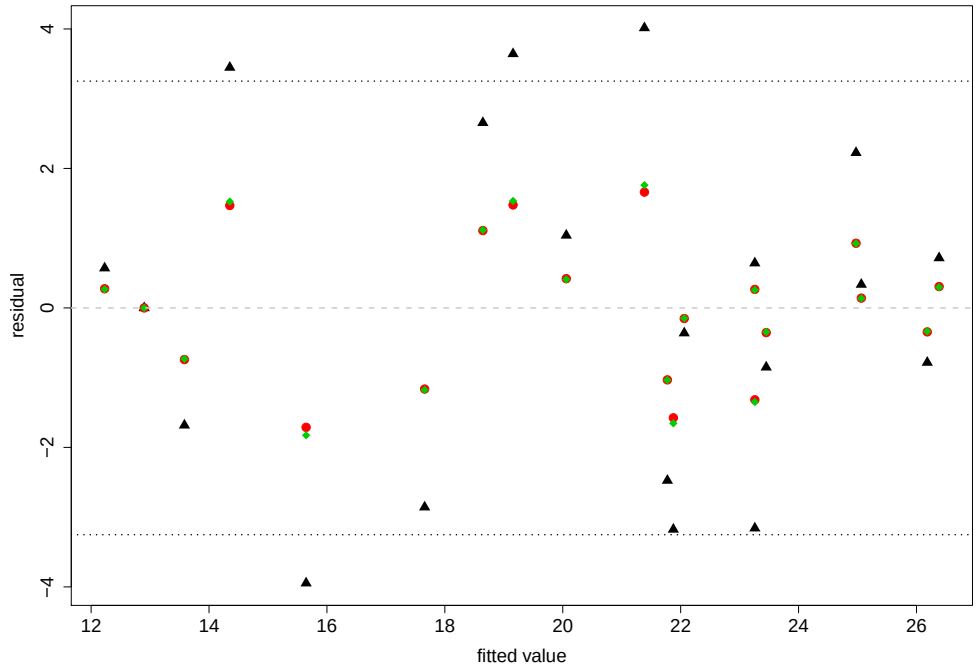
Body Fat: Model 3 $Y \sim X_1, X_2$

	h_ii	e_i	r_i	d_i	t_i
1	0.201	-1.683	-0.740	-2.106	-0.730
2	0.059	3.643	1.477	3.871	1.534
3*	0.372	-3.176	-1.576	-5.057	-1.654
4	0.111	-3.158	-1.317	-3.553	-1.348
5	0.248	0.000	0.000	0.000	0.000
6	0.129	-0.361	-0.152	-0.414	-0.148
7	0.156	0.716	0.306	0.848	0.298
8*	0.096	4.015	1.661	4.442	1.760
9	0.115	2.655	1.110	2.999	1.118
10	0.110	-2.475	-1.032	-2.781	-1.034
11	0.120	0.336	0.141	0.382	0.137
12	0.109	2.226	0.927	2.499	0.923
13*	0.178	-3.947	-1.712	-4.804	-1.826
14	0.148	3.447	1.469	4.046	1.525
15	0.333	0.571	0.275	0.856	0.267
16	0.095	0.642	0.266	0.710	0.258
17	0.106	-0.851	-0.354	-0.951	-0.345
18	0.197	-0.783	-0.344	-0.975	-0.334
19	0.067	-2.857	-1.163	-3.062	-1.176
20	0.050	1.040	0.420	1.095	0.409

Cases with top three largest studentized deleted residuals are marked by *.

For $\alpha = 0.1$, $t(1 - \alpha/40; 20 - 3 - 1) = 3.25$, so there is no significant outliers. We may still want to investigate the top cases, since Bonferroni's procedure is very conservative (and thus lack of power).

Figure: Body Fat: Residuals vs. fitted values plot. Black – ordinary residuals, Red – studentized residuals, Green – studentized deleted residuals. Black dotted lines ($h = \pm 3.25$) correspond to Bonferroni's procedure critical value at $\alpha = 0.1$.



No obvious outliers.

```
n=nrow(fat) ## number of cases
p=3 ## number of parameters

fit3=lm(Y~X1+X2, data=fat) ## Model 3
MSE=anova(fit3)["Residuals",3] ## MSE of Model 3 fit
res=fit3$residuals ## residuals
hh = influence(fit3)$hat ## diagonal of the hat matrix: leverage values

stu.res=res/sqrt(MSE*(1-hh)) ## studentized residuals

res.del=res/(1-hh) ##deleted residuals
library(MASS)
stu.res.del=studres(fit3) ## studentized deleted residuals
alpha=0.1
bon.thre=qt(1-alpah/(2*n), n-p-1) ## Bonferroni's threshold at alpha

## residuals vs. fitted values plots
plot(fit3$fitted, res, xlab="fitted value", ylab="residual",
+ cex.lab=1.5, cex.axis=1.5, pch=17, cex=1.5)
points(fit3$fitted, stu.res, col=2, pch=19, cex=1.5)
points(fit3$fitted, stu.res.del, col=3, pch=18, cex=1.5)
abline(h=0, col=grey(0.8), lwd=2, lty=2)
abline(h=bon.thre, lwd=2, lty=3)
abline(h=-bon.thre, lwd=2, lty=3)
```

Leverage Values

The *ith* diagonal element h_{ii} of the hat matrix **H** is called the *leverage* of the *ith* case.

- The fitted value $\widehat{Y}_i$ is a linear combination of the observations:

$$\widehat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j.$$

  - Note $h_{ii} + \sum_{j \neq i} h_{ij} = 1$: So the larger h_{ii} is, the more important Y_i is in determining $\widehat{Y}_i$.
  - The hat matrix only depends on the design matrix. So h_{ii} **measures the role of the X values in the importance of Y_i in terms of determining the fitted value $\widehat{Y}_i$.**

Identify Outlying X by Leverage

$$h_{ii} = \frac{1}{n} + \mathbf{x}_i^{*T}(\mathbf{r}_{XX})^{-1}\mathbf{x}_i^*, \quad \mathbf{x}_i^{*T} = \frac{1}{\sqrt{n-1}}(X_{i1} - \overline{X}_1, \cdots, X_{i,p-1} - \overline{X}_{p-1}).$$

- h_{ii} reflects the *Mahalanobis distance* between the X values of the *ith* case $\begin{bmatrix} X_{i1} & X_{i2} & \cdots X_{i,p-1} \end{bmatrix}^T$ and the sample mean of the X values (center of X): $\overline{\mathbf{x}} = \begin{bmatrix} \overline{X}_1 & \overline{X}_2 & \cdots & \overline{X}_{p-1} \end{bmatrix}^T$.
  - h_{ii} will achieve the minimum value $1/n$ when the X values of the *ith* case coincide with the center of X, i.e., when $\mathbf{x}_i^* = \mathbf{0}_{p-1}$.
  - A large value of h_{ii} indicates that the X values of the *ith* case is far away from the center of X when taking into account of the shape of the observed data cloud.
  - **A large leverage value is an indication of potential outlying in X.**

Geometric Interpretation of Leverage

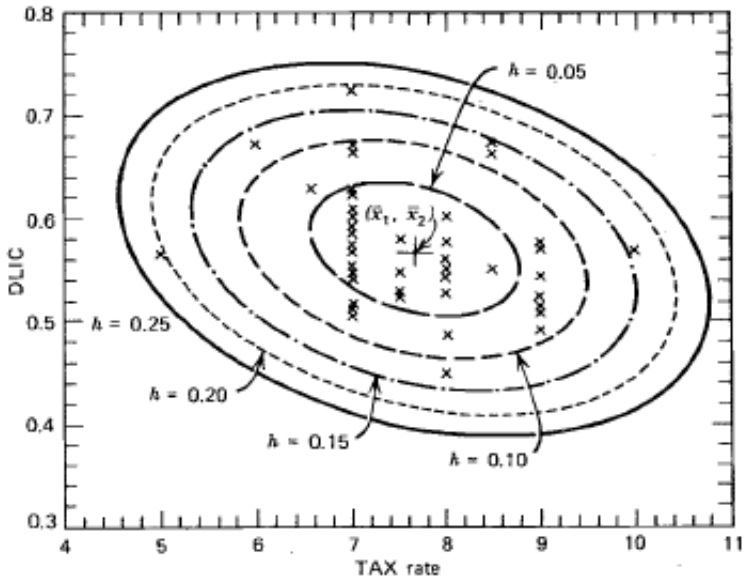


Figure 5.2 Contours of constant h_{ii} in two dimensions. [1]

The term $\mathbf{x}_i^{*T}(\mathbf{r}_{XX})^{-1}\mathbf{x}_i^*$ gives the equation of an ellipsoid centered at $\bar{\mathbf{x}}$. Having the same Euclidean distance from $\bar{\mathbf{x}}$, points along the major direction of the data cloud have values of h_{ii} than points along the minor direction of the data cloud. In this sense, points along the major direction are the center $\bar{\mathbf{x}}$.

[1] Figure from S. Weisberg, Applied linear regression.

Geometric Interpretation of Leverage

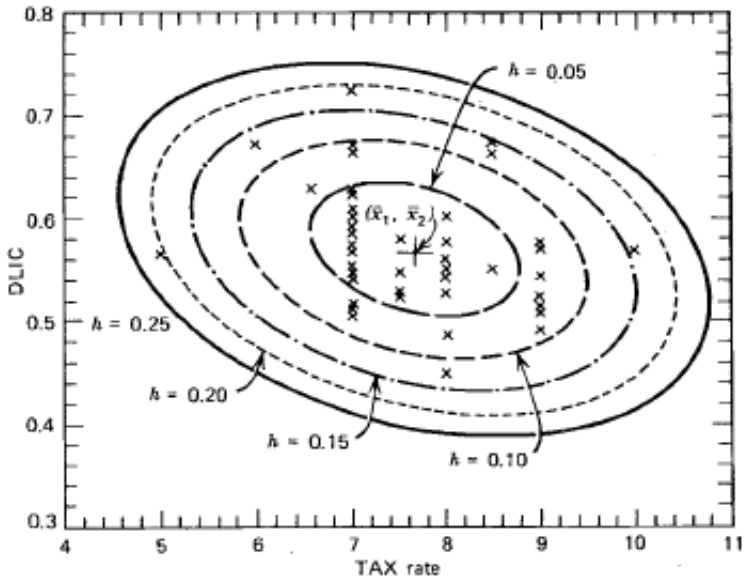


Figure 5.2 Contours of constant h_{ii} in two dimensions. [2]

The term $\mathbf{x}_i^{*T}(\mathbf{r}_{XX})^{-1}\mathbf{x}_i^*$ gives the equation of an ellipsoid centered at $\overline{\mathbf{x}}$. Having the same Euclidean distance from $\overline{\mathbf{x}}$, points along the major direction of the data cloud have smaller values of h_{ii} than points along the minor direction of the data cloud. In this sense, points along the major direction are closer to the center $\overline{\mathbf{x}}$.

[2] Figure from S. Weisberg, Applied linear regression.

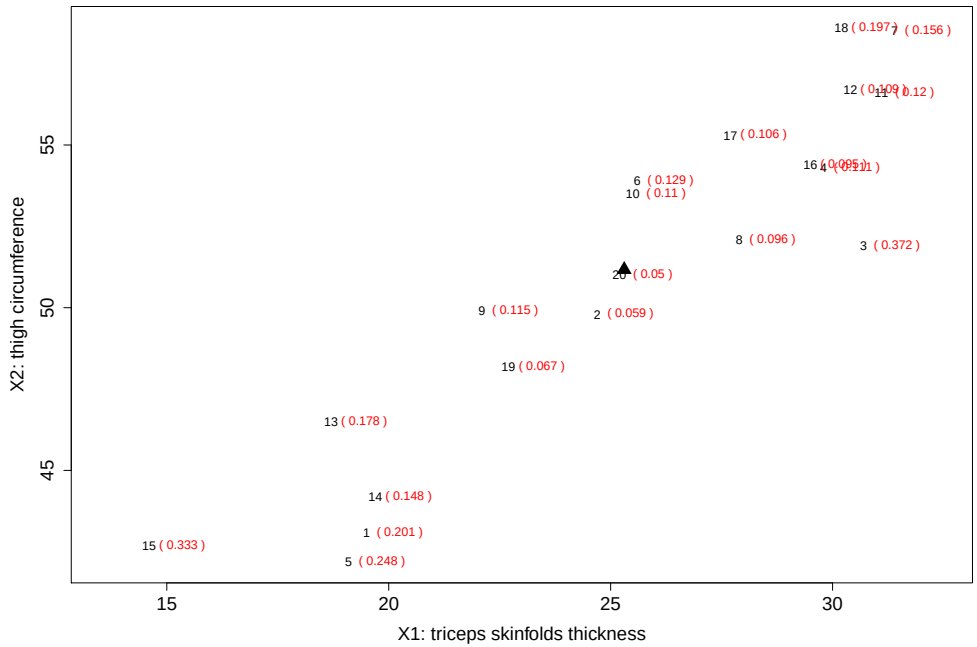
In practice, a leverage value is often considered to be large if it is more than twice as large as the mean leverage value $\bar{h}$:

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

- If $h_{ii} > \frac{2p}{n}$, then the *ith* case is identified as outlying with regard to its X values.
- The above rule is only applicable when the sample size n is relatively large compared to the number of parameters p.

Body Fat: Model 3 Leverage Values

Figure: Body Fat: Scatter plot of X_2 vs. X_1. Data points are identified by case numbers. Numbers in parenthesis are leverage values. Black triangle is the center of X values.



Here $n = 20, p = 3, \frac{2p}{n} = 0.3$. Two cases, 15 and 3, have leverage values greater than 0.3.

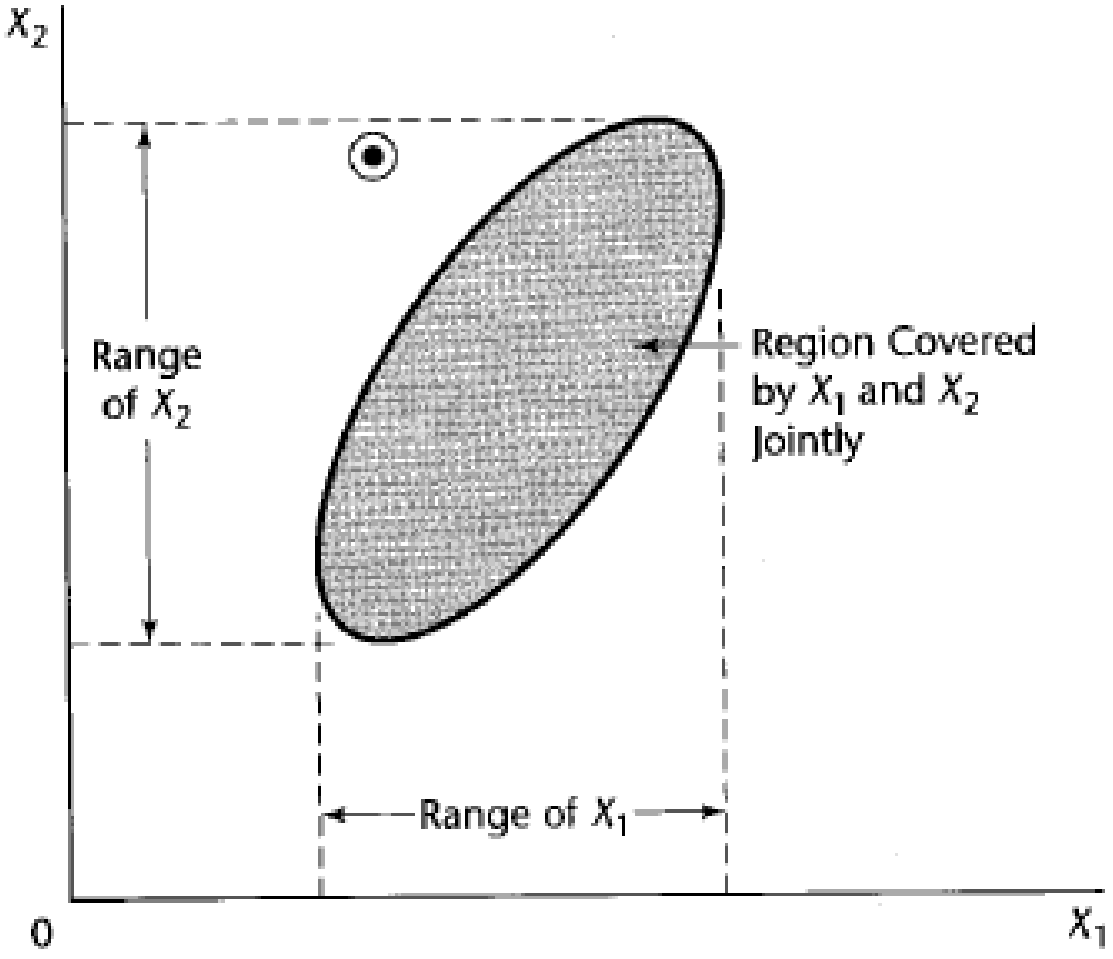
- Case 15 is outlying in terms of and is at the low end of the range for X_2. $h_{15,15} = 0.333$.
- Case 3 is outlying in terms of

  , though it is for either X_1 or X_2 individually. $h_{33} = 0.372$.
- The third and fourth largest leverage values are $h_{55} = 0.248$ and $h_{11} = 0.201$ which are substantially smaller than h_{33} and $h_{15,15}$. From the plot, cases 1 and 5 are somewhat outlying.
- The next task is to determine how influential cases 3 and 15 are on the fitted regression function.

- Case 15 is outlying in terms of X_1 and is at the low end of the range for X_2. $h_{15,15} = 0.333$.
- Case 3 is outlying in terms of the pattern of association between X_1 and X_2, though it is not outlying for either X_1 or X_2 individually. $h_{33} = 0.372$.
- The third and fourth largest leverage values are $h_{55} = 0.248$ and $h_{11} = 0.201$ which are substantially smaller than h_{33} and $h_{15,15}$. From the plot, cases 1 and 5 are somewhat outlying.
- The next task is to determine how influential cases 3 and 15 are on the fitted regression function.

Hidden Extrapolations

- Extrapolation occurs when predicting the response variable/estimating the mean response for X values lying outside the range of the X in the data set used to fit the model.
- The fitted model may not be appropriate when extended outside the range of the original X.
- With more than one X variables, **the levels of all X variables jointly define the region of the observations.** One can not merely look at the range of each X variable separately.
- With more than two X variables, we can utilize the leverage calculation to identify extrapolation.


X_2
Range of X_2
Region Covered by X_1 and X_2 Jointly
Range of X_1
0
X_1

Identify Hidden Extrapolation by Leverage

Leverage calculation for the new X for which prediction/mean response estimation is to be made:

$$h_{new,new} = \mathbf{x}_{new}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}.$$

- $\mathbf{x}_{new}$ is the column vector containing the new X and $\mathbf{X}$ is the design matrix of the data used for fitting the regression model.
- If $h_{new,new}$ is within the range of leverage values h_{ii} for cases in the data set, then no extrapolation occurs.
- If $h_{new,new}$ is much greater than the leverage values h_{ii} for cases in the data set, then an extrapolation is indicated.

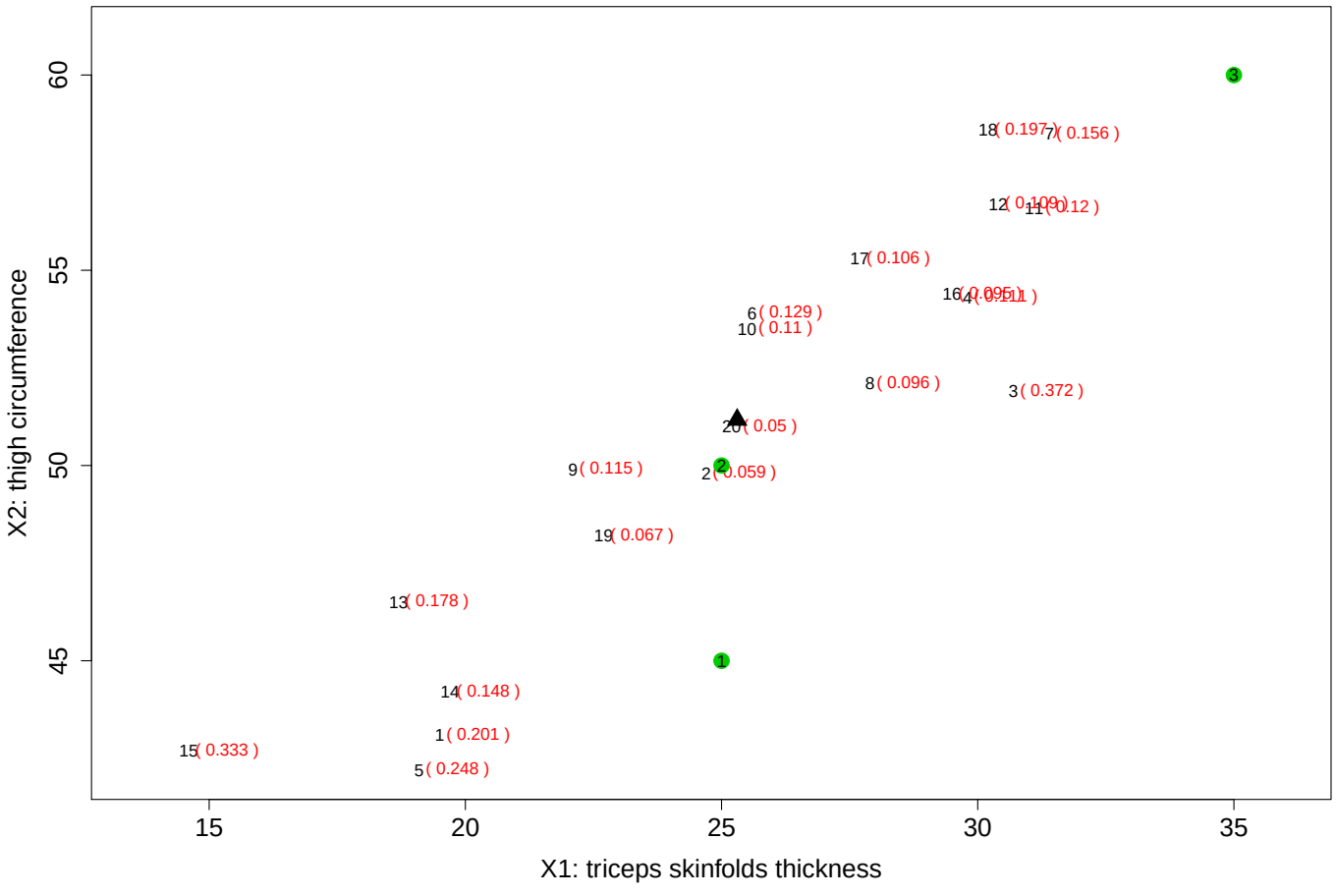
Body Fat: Hidden Extrapolations

```
> range(fat[,1]) ## range of X1
[1] 14.6 31.4
> range(fat[,2]) ##range of X2
[1] 42.2 58.6
> range(hh)## range of leverage values
[1] 0.05008526 0.37193301
>
> xnew1=c(1,25, 45) ## within both ranges of X1 and X2
> hnew1=t(xnew1)%*%solve(t(X)%*%X)%*%xnew1
> hnew1  ## hidden extrapolation since not consistent with the pattern
          [,1]
[1,] 0.5028977

> xnew2=c(1,25, 50) ## within both ranges of X1 and X2
> hnew2=t(xnew2)%*%solve(t(X)%*%X)%*%xnew2
> hnew2    ## no extrapolation
           [,1]
[1,] 0.06026272

> xnew3=c(1,35, 60) ## somewhat outside of ranges
> hnew3=t(xnew3)%*%solve(t(X)%*%X)%*%xnew3
> hnew3   ## no  extrapolation since consistent with the pattern
          [,1]
[1,] 0.2493753
```

Figure: Body Fat: Hidden Extrapolations



Identify Influential Cases

We want to determine whether the outlying cases (in Y and/or in X) are influential in regression function fitting.

- A case is considered to be *influential* if its exclusion leads to major changes of the fitted regression function.
- Cook's distance.
  - It measures the aggregate influence on all fitted values that is made by the omission of a single case in the fitting process.

Cook's Distance

Cook's distance measures the aggregate influence of the *ith* case on all n fitted values:

$$D_i := \frac{\sum_{j=1}^{n}(\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{p \times MSE}, \quad i = 1, \cdots, n.$$

- $\widehat{Y}_j$ is the fitted value for the *jth* case when all cases are used to derive the fitted regression function.
- $\widehat{Y}_{j(i)}$ is the fitted value for the *jth* case when the *ith* case is excluded from the fitting process.
- $p \times MSE$ serves as a standardizing quantity.
- Calculate $p_i = P(F_{p,n-p} < D_i)$.
  - If $p_i < 10\% \sim 20\%$, then D_i is deemed to be small and the *ith* case has little influence on the fitted values.
  - If $p_i > 50\%$, then D_i is deemed to be large and the *ith* case has a large influence on the fitted values.

Cook's distance can also be computed from the regression fit based on all cases:

$$D_i = \frac{e_i^2}{p \times MSE} \frac{h_{ii}}{(1-h_{ii})^2} = \frac{r_i^2}{p} \frac{h_{ii}}{(1-h_{ii})}, \quad (1)$$

where $r_i = e_i / \sqrt{MSE(1-h_{ii})}$ is the ith studentized residual.

- The magnitude of D_i depends on two factors (i) the studentized residual r_i; and (ii) the leverage value h_{ii}. The larger $|r_i|$ and/or h_{ii} is, the D_i tends to be.
- So an influential case could be due to either outlying in Y (a large studentized residual) or outlying in X (a large leverage value) or both.
- On the other hand, outlying in Y or outlying in X alone

  a case influential.

Cook's distance can also be computed from the regression fit based on all cases:

$$D_i = \frac{e_i^2}{p \times MSE} \frac{h_{ii}}{(1 - h_{ii})^2} = \frac{r_i^2}{p} \frac{h_{ii}}{(1 - h_{ii})}, \tag{2}$$

where $r_i = e_i / \sqrt{MSE(1 - h_{ii})}$ is the ith studentized residual.

- The magnitude of D_i depends on two factors (i) the size of the studentized residual r_i; and (ii) the leverage value h_{ii}. The larger $|r_i|$ and/or h_{ii} is, the larger D_i tends to be.
- So an influential case could be due to either outlying in Y (a large studentized residual) or outlying in X (a large leverage value) or both.
- On the other hand, outlying in Y or outlying in X **alone** does not necessarily make a case influential.

Body Fat: Cook's Distance

- Consider Cook's distance for case 3. It has a residual $e_3 = -3.176$ and leverage value $h_{33} = 0.372$. Also $p = 3$ and $MSE = 6.47$. So

$$D_3 = \frac{(-3.176)^2}{3 \times 6.47} \frac{0.372}{(1 - 0.372)^2} = 0.49.$$

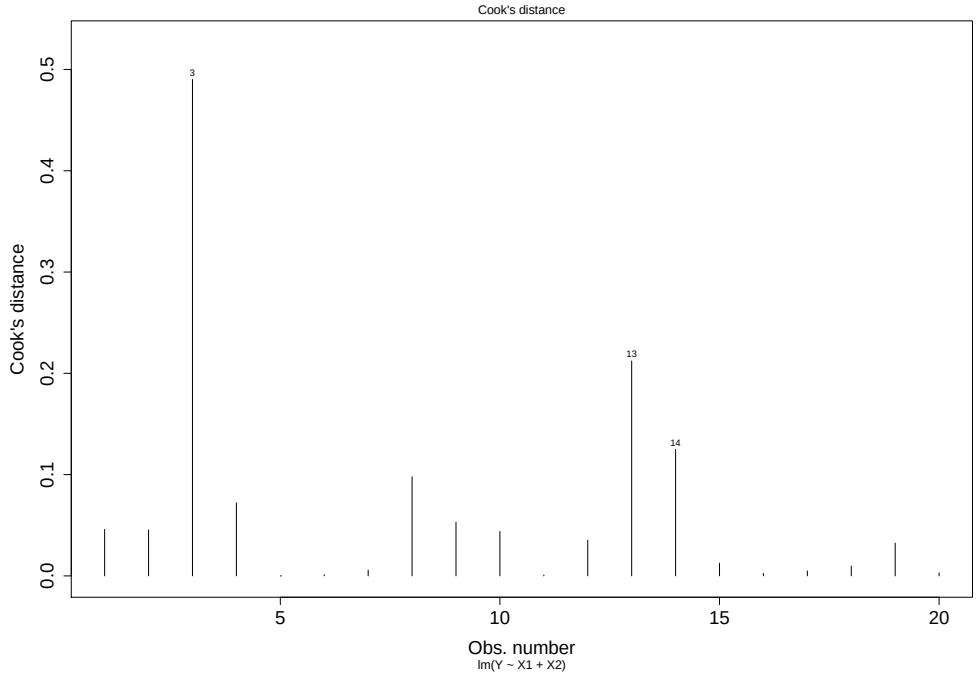
- To assess the magnitude of D_3, we compare it with the F distribution $F_{3,17}$ ($p = 3, n - p = 20 - 3 = 17$):

$$p_3 = P(F_{3,17} < 0.49) = 30\%.$$

- Therefore, case 3 has some aggregated influence on all the fitted values, but the influence does not seem to be large enough to call for remedial measures.

Cook's Distance: Index Influence Plot

Figure: Body Fat:Cook's distance vs. case index



Case 3 stands out as much more influential than other cases according to Cook's distance measure.

```
plot(fit3, which=4) ## cook's distance
```

Influential Cases: Comments

It is often a good idea to directly examine the influences on the fitted regression function with and without the case(s) of concern.

- If inferences of interest (e.g., prediction) are not much affected, then there is no need to call for remedial measures.
- If serious changes occur, then we need to think about remedial measures, e.g., drop the case(s) of concern or consider *robust regression*.