Stat 206: Linear Models

Lecture 16

Nov. 25, 2019

Model Validation

- *Internal validation*: Check validity using **the same data** used to fit the model.
- *External validation*: Check validity using **new data** – either newly collected or a holdout sample.
- Compare results with theoretical expectations, previous results, and simulation results.

Internal Validation

- $Press_p$ **is always** **than** SSE_p as

$$|d_i| = |Y_i - \widehat{Y}_{i(i)}| = |\frac{Y_i - \widehat{Y}_i}{1 - h_{ii}}| \geq |Y_i - \widehat{Y}_i| = |e_i|, \quad i = 1, \cdots, n.$$

  - $Press_p/n$ can be viewed as an estimator of the *(out-of-sample) mean squared prediction error*:

$$mspe := E((\hat{y} - y)^2).$$

  - It is a measure of the of the model.
  - $Press_p$ not much larger than SSE_p means there is by the model.
- $C_p \approx p$ indicates in the model, whereas $C_p >> p$ indicates model bias.

Training Data vs. Validation Data

When sample size is sufficiently large, we can split the data into two sets, a *training data* used to build the model and a *validation data* used to check model validity.

- Validation data is used to check consistency of the fitted parameters and predictive ability.
- Training data should be sufficiently large (e.g., n/P at least 6) so that a reliable model can be built based on it. Sometimes, the validation data will have to be smaller.
- Once a final model has been validated and chosen, it is a common practice to use the entire data set to re-fit the final model.

Mean Squared Prediction Error

$$MSPE_v = \frac{\sum_{j=1}^{m}(Y_j - \widehat{Y}_j)^2}{m}.$$

m is the sample size of the validation data, Y_j is the *jth* observation in the validation data, and $\widehat{Y}_j$ is the predicted value of the *jth* case based on the model fitted on the training data.

- $MSPE_v$ can be viewed as an estimator of the *(out-of-sample) mean squared prediction error* and thus a measure for the predictive ability of the model.
- $MSPE_v$ is usually than SSE/n, since the model is fitted on the training data and thus it naturally would fit the training data than it fits the validation data.
- If $MSPE_v$ is not much larger than SSE/n, then there is by the model.

Surgical Unit: Internal Validation

Three “best” models according to various criteria.

- By BIC_p and $Press_p$: Model 1, $\log Y \sim X_1, X_2, X_3, X_8$.
  - $p = 5, SSE_p = 2.178, C_p = 5.734, Press_p = 2.736$.
- By C_p: Model 2, $\log Y \sim X_1, X_2, X_3, X_6, X_8$.
  - $p = 6, SSE_p = 2.081, C_p = 5.528, Press_p = 2.782$.
- By $R^2_{a,p}$ and AIC_p: Model 3, $\log Y \sim X_1, X_2, X_3, X_5, X_6, X_8$.
  - $p = 7, SSE_p = 2.004, C_p = 5.772, Press_p = 2.771$.
- For all three models, $Press_p$ and SSE_p are reasonably close and $C_p \approx p$, supporting their validity.

Surgical Unit: External Validation

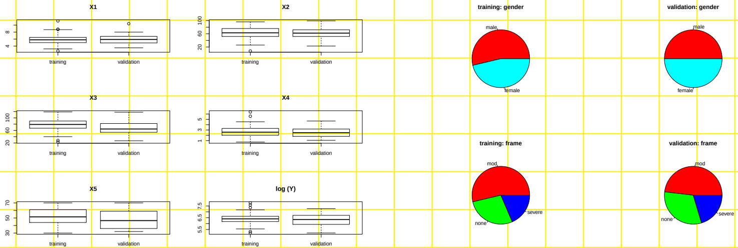


Figure: Distributions of variables in training ($n = 54$) and validation ($n = 54$) sets.

No big difference in how variables are distributed in these two sets.

All three models have:

- Consistency in parameter estimation: same sign and similar magnitude between the two sets of estimated coefficients and their standard errors.
- $MSPE_v$ based on the validation data is not much larger than SSE/n and Press/n based on the training data.

Surgical Unit: Model 1 External Validation

```
## fit model 1 on training and validation sets
> fit1=lm(log(Y)~X1+X2+X3+X8, data=data.o)
> fit1.v=lm(log(Y)~X1+X2+X3+X8, data=data.v)

## get estimates and statistics
> est1=cbind(summary(fit1)$coefficients[,1:2], summary(fit1.v)$coefficients[,1:2])
> sse1=c(anova(fit1)["Residuals",2],anova(fit1.v)["Residuals",2])
> mse1=c(anova(fit1)["Residuals",3],anova(fit1.v)["Residuals",3])
> Rs1=c(summary(fit1)$adj.r.squared, summary(fit1.v)$adj.r.squared)
> press1= c(sum(fit1$residuals^2/(1-influence(fit1)$hat)^2),
+ sum(fit1.v$residuals^2/(1-influence(fit1.v)$hat)^2))

## MSPE on validation set
> newdata=data.v[,1:8] ## validation cases for prediction
> mspe1=c('NA', mean((predict.lm(fit1, newdata)-log(data.v$Y))^2))

## display results
> temp=cbind(sse1, mse1, Rs1, press1, press1/n,mspr1)
> rownames(temp)=c("Training", "Validation")
> colnames(temp)=c("sse","mse","R2_a","press","press/n", "mspe")
> round(est1,3)
> round(temp,3)
```

Surgical Unit: Model 1 External Validation (Cont'd)

	Training		Validation	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	3.853	0.193	3.635	0.289
X1	0.073	0.019	0.096	0.032
X2	0.014	0.002	0.016	0.002
X3	0.015	0.001	0.016	0.002
X8	0.353	0.077	0.186	0.096

	sse	mse	R2_a	press	press/n	mspe
Training	2.178	0.044	0.816	2.736	0.051	--
Validation	3.794	0.077	0.682	--	--	0.077

Surgical Unit: Model 2 External Validation

	Training		Validation	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	3.867	0.191	3.614	0.291
X1	0.071	0.019	0.100	0.032
X2	0.014	0.002	0.016	0.002
X3	0.015	0.001	0.015	0.002
X6	0.087	0.058	0.073	0.079
X8	0.363	0.077	0.189	0.097

	sse	mse	R2_a	press	press/n	mspe
Training	2.081	0.043	0.821	2.782	0.052	--
Validation	3.728	0.078	0.682	--	--	0.076

Surgical Unit: Model 3 External Validation

```
Training                Validation
Estimate Std. Error Estimate Std. Error
(Intercept)     4.054       0.235     3.470       0.347
X1              0.072       0.019     0.099       0.032
X2              0.014       0.002     0.016       0.002
X3              0.015       0.001     0.016       0.002
X5             -0.003       0.003     0.003       0.003
X6              0.087       0.058     0.073       0.079
X8              0.351       0.076     0.193       0.097
```

```
sse   mse   R2_a press press/n  mspe
Training   2.004 0.043 0.823 2.771    0.051   --
Validation 3.681 0.078 0.679  --       --    0.079
```

Surgical Unit: Choice of Final Model

- $MSPE_v$ of the three models have similar values, indicating that they have similar predictive ability.
- Model 3 has one estimated regression coefficient changing sign from training data to validation data, probably due to relatively large SE of this coefficient. So it is eliminated from further consideration.
- Models 1 and 2 perform similarly in validation. Based on the **principle of parsimony**, we choose Model 1 as the final model.
- Fit Model 1 on all data ($n = 108$):

$$\begin{aligned} \log \quad & (\text{Survial Time}) = 3.76 + 0.084 \times \text{clotting score} \\ + \quad & 0.015 \times \text{ prognostic index} + 0.016 \times \text{enzyme score} \\ + \quad & 0.265 \times I(\text{severe use of alcohol}). \end{aligned}$$

Surgical Unit: Final Model Fitted on All Data

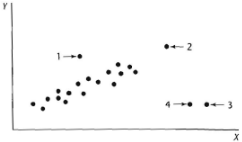
```
Call:
lm(formula = log(Y) ~ X1 + X2 + X3 + X8, data = rbind(data.o,
data.v))
Residuals:
Min        1Q   Median       3Q      Max
-0.60369 -0.15201  0.00977  0.13175  0.57726
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.756276   0.162825  23.069  < 2e-16 ***
X1          0.083744   0.016781   4.990 2.46e-06 ***
X2          0.014988   0.001409  10.641  < 2e-16 ***
X3          0.015690   0.001134  13.839  < 2e-16 ***
X8          0.265096   0.060045   4.415 2.50e-05 ***
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
Residual standard error: 0.2446 on 103 degrees of freedom
Multiple R-squared: 0.7642,     Adjusted R-squared: 0.755
F-statistic: 83.45 on 4 and 103 DF,  p-value: < 2.2e-16
> anova(fit1.all)
Analysis of Variance Table

Response: log(Y)
Df   Sum Sq Mean Sq F value    Pr(>F)
X1          1  1.0809  1.0809  18.064 4.703e-05 ***
X2          1  6.5415  6.5415 109.322 < 2.2e-16 ***
X3          1 11.1859 11.1859 186.940 < 2.2e-16 ***
X8          1  1.1663  1.1663  19.492 2.498e-05 ***
Residuals 103  6.1632  0.0598
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
```

Outlying Cases

Data may contain cases that are outlying or extreme:

- 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. These are called *influential cases*.
- It is important to identify outlying cases and to investigate their effects in order to decide whether they should be retained or eliminated.

Examples of Outlying Cases



- Case 1 is outlying in , but influential: There are a few other cases with X values and they will keep the fitted regression function from being distorted too much by Case 1.
- Case 2 is outlying in , but influential: Its Y value is with the regression relation suggested by other cases.
- Cases 3 and 4 are influential: They are outlying in X and their Y values are with the regression relation suggested by other cases.

Identify Outlying Cases

- With one or two X variables, outlying cases can be identified by graphs 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.
- 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

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}.$$

- Assume $Var(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$, then
  $\sigma^2\{\mathbf{e}\} =$, $\mathbf{s}^2\{\mathbf{e}\} =$.
- 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.

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 one another.
- In the R function *plot.lm()*, the residuals QQ plot (which=2), scale-location plot (which=3) and residuals vs. leverage plot (which=5) use studentized residuals.

Deleted Residuals

To be more effective in detecting outlying Y, 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 could be by the *ith* case to be to Y_i such that e_i may 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 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).

Studentized Deleted Residuals

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

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

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

Under H_0: The model is correct and all cases follow the model

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

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.

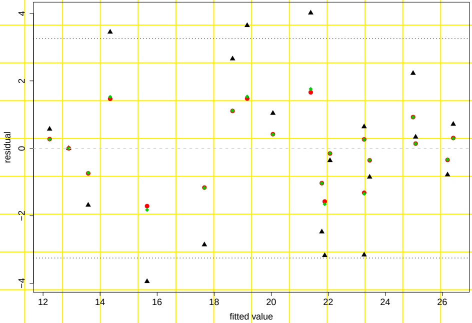
Body Fat: Model 3 $Y \sim X_1, X_2$

	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. However, we may still want to investigate the top few cases.

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 $\mathbf{H}$ is called the *leverage* of the *ith* case.

- The fitted value $\widehat{Y}_i$:

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

  - Recall $h_{ii} + \sum_{j \neq i} h_{ij} = 1$ and $1/n \leq h_{ii} \leq 1$: The larger h_{ii} is, the more important Y_i is in determining $\widehat{Y}_i$.
  - h_{ii} measures the role of the X values in terms of determining the fitted value $\widehat{Y}_i$.

Identify Outlying X by Leverage

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

- 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): $\bar{\mathbf{x}} = \begin{bmatrix} \overline{X}_1 & \overline{X}_2 & \cdots & \overline{X}_{p-1} \end{bmatrix}^T$.
  - A large value of h_{ii} indicates that the X values of the *ith* case is the center of X when taking into account of the shape of the observed data cloud.
  - A large leverage value is an indication of

Geometric Interpretation of Leverage

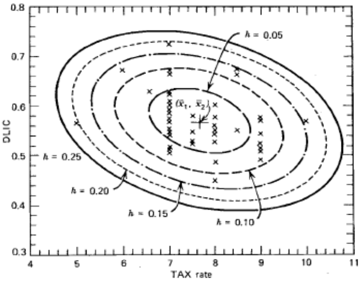


Figure 5.2 Contours of constant h_{ii} in two dimensions.

From S. Weisberg, Applied linear regression

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 $\bar{\mathbf{x}}$.

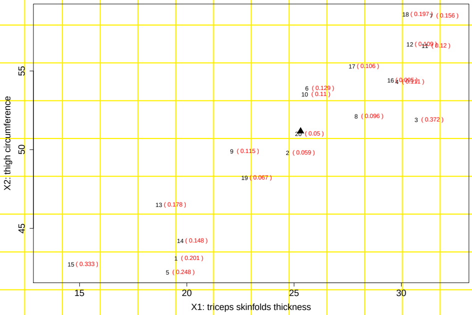
In practice, a leverage value is often considered to be large if it is more than twice as large as the mean leverage value $\bar{h}$:

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

- 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 not too small.

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.



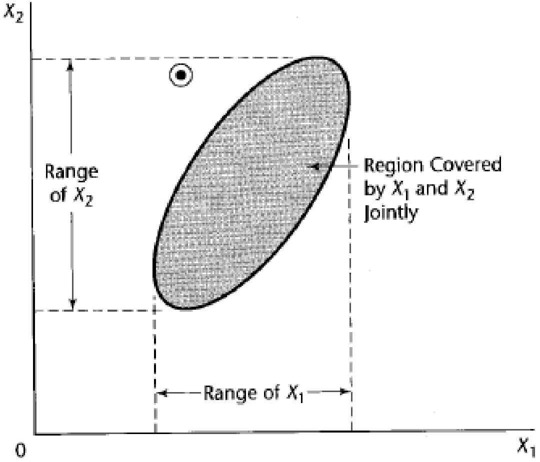
Here $n = 20, p = 3, \frac{2p}{n} = 0.3$. Two cases, 15 and 3, have leverage values greater than 0.3.

- 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.

Hidden Extrapolations

- Extrapolation occurs when the response variable/estimating the mean response for X values the region of the X in the data used to fit the model.
- With more than one X variables, the levels of the range of the observations. One can not merely look at the region 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

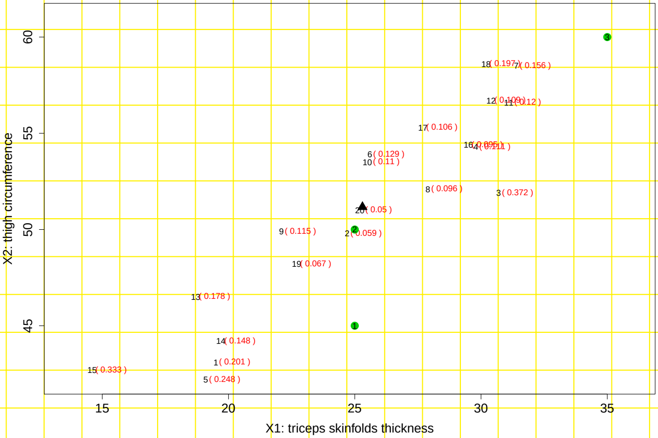
Leverage calculation for a new X:

- $\mathbf{x}_{new}$ is the column vector containing the new X and $\mathbf{X}$ is the design matrix of the data used to fit the regression model.
- If $h_{new,new}$ is of leverage values h_{ii} for cases in the data set, then no extrapolation occurs.
- If $h_{new,new}$ is the leverage values h_{ii}, 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 determining the fitted regression function.

- 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

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

- $\widehat{Y}_j$ 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 standardization quantity.
- In practice, $D_i > \frac{4}{n-p}$ is often used as an indicator for being a potential influential case.
- A more conservative approach is to use $D_i > 1$ as the cutoff for influential cases.

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

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

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

- If case i follows the same regression relation as other cases, then $E(D_i) \approx \frac{h_{ii}}{p(1-h_{ii})} \sim \frac{1}{n-p}$ when n is large (as $h_{ii} \sim p/n$ and $E(r_i^2) \approx 1$)).
- 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 or or both.
- On the other hand, outlying in Y or outlying in X alone

  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 $\frac{4}{n-p} = \frac{4}{20-3} = 0.23$.
- Therefore, case 3 has some aggregated influence on all the fitted values and may need further investigation.

Cook's Distance: Index Influence Plot

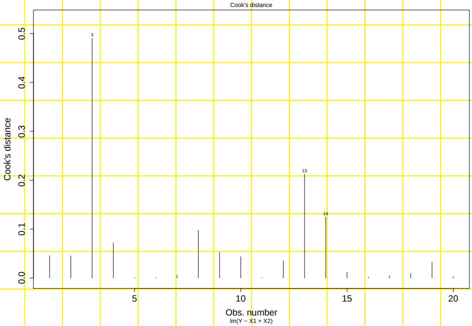


Figure: Body Fat:Cook's distance vs. case index

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

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