Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

Model Selection: Cont'd

Recap: Full Model vs. Candidate Model

- *Full model*: The model that contains all $P - 1$ potential X variables in the pool.
  - **Assume the full model is a correct model**.
- *Candidate model*: A model that contains a subset of $p - 1$ X variables with $1 \leq p \leq P$.
- The goal is to choose good model(s) (subset(s) of X variables) that balances bias and variance.

Recap: Key Components for Model Selection

- **Criterion to compare models**:
  - $R_a^2, C_p, AIC_p, BIC_p, Press_p$, etc.
- **Procedure to search for good model(s):**
  - *Best subset selection*: Exhaustive search; Applicable when the number of potential X variables is not too big ;
  - *Stepwise regression*: Greedy search; The number of potential X variables can be large;

Surgical Unit

capital $P = 5$

Cp: want to choose the one not too much larger than P

If clotting (X_1), prognostic (X_2), enzyme (X_3), liver (X_4) form the potential pool of X variables, then there are 16 sub-models.

always included

2^4

p	intercept	X1	X2	X3	X4	sse	R^2	R^2_a	Cp	aic	bic	press
1	1	0	0	0	0	12.805	0.000	0.000	151.569	-75.716	-73.727	13.292
2	1	0	0	1	0	7.334	0.427	0.416	66.518	-103.811	-99.833	8.329
2	1	0	0	0	1	7.408	0.421	0.410	67.696	-103.268	-99.290	8.024
2	1	0	1	0	0	9.974	0.221	0.206	108.469	-87.205	-83.227	10.738
2	1	1	0	0	0	12.028	0.061	0.043	141.093	-77.096	-73.118	13.508
3	1	0	1	1	0	4.313	0.663	0.650	20.523	-130.479	-124.512	5.066
3	1	0	0	1	1	5.132	0.599	0.583	33.536	-121.089	-115.122	6.123
3	1	1	0	1	0	5.783	0.548	0.531	43.873	-114.644	-108.677	6.989
3	1	0	1	0	1	6.620	0.483	0.463	57.175	-107.342	-101.375	7.474
3	1	1	0	0	1	7.299	0.430	0.408	67.961	-102.070	-96.103	8.472
3	1	1	1	0	0	9.437	0.263	0.234	101.937	-88.194	-82.227	11.055
4	1	1	1	1	0	3.109	0.757	0.743*	3.388*	-146.161*	-138.205*	3.914*
4	1	0	1	1	1	3.615	0.718	0.701	11.434	-138.011	-130.055	4.598
4	1	1	0	1	1	4.970	0.612	0.589	32.960	-120.823	-112.867	6.209
4	1	1	1	0	1	6.568	0.487	0.456	58.358	-105.763	-97.807	7.902
5	1	1	1	1	1	3.084	0.759*	0.739	5.000	-144.587	-134.642	4.069

"none model"

full model

always minimize

Here we can do exhaustive subset ?

(< 21)

all criterion choose this (not always)

Model Selection: Criteria (Cont'd)

Recap: Mallows' C_p Criterion

$$C_p := \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p)$$

- n : sample size
- p: number of regression coefficients in the candidate model
- SSE_p: error sum of squares of the candidate model
- $\hat{\sigma}^2 = MSE_{\text{full model}}$
- Look for models with (i) the C_p value not far above p and (ii)

  less X variables $\Longrightarrow$ small bias and small variance

Recap: AIC_p and BIC_p Criteria

- *Akaike's information criterion (AIC)*:

If bias is top concern

$$AIC_p = n \log \frac{SSE_p}{n} + 2p$$

- *Bayesian information criterion (BIC)*:

Smaller model

$$BIC_p = n \log \frac{SSE_p}{n} + (\log n)p$$

more penalty on complexity

- Look for models with small AIC (BIC)

$Press_p$ Criterion

Predicted residual sum of squares ($Press_p$):

$$Press_p = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{i(i)})^2.$$

- Y_i is the observed response of the *ith* case.
- $\widehat{Y}_{i(i)}$ is the predicted value for the ith case obtained by fitting the model only using $n - 1$ cases excluding case i.

- $Press_p$ is also known as *leave-one-out-cross-validation (LOOCV)*.

- Models with small $Press_p$ are considered good in terms of predictive ability.

$Press_p$: Calculation

$Press_p$ can be calculated without actually performing n regressions because the *deleted residual* for the *ith* case:

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

where $e_i = Y_i - \widehat{Y}_i$ is the residual of the *ith* case and h_{ii} is the *ith* diagonal element of the hat matrix $\mathbf{H}$, both from the regression fit using **all** n cases. So

$$Press_p = \sum_{i=1}^{n} \frac{(Y_i - \widehat{Y}_i)^2}{(1 - h_{ii})^2}.$$

Surgical Unit: Full Model

```
lm(formula = log(Y) ~ X1 + X2 + X3 + X4, data = data.o)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.851933   0.266263  14.467  < 2e-16 ***

X1          0.083739   0.028834   2.904  0.00551 **

X2          0.012671   0.002315   5.474 1.50e-06 ***

X3          0.015627   0.002100   7.440 1.38e-09 ***

X4          0.032056   0.051466   0.623  0.53627

Residual standard error: 0.2509 on 49 degrees of freedom

Multiple R-squared: 0.7591,     Adjusted R-squared: 0.7395

F-statistic: 38.61 on 4 and 49 DF,  p-value: 1.398e-14


Analysis of Variance Table

Df Sum Sq Mean Sq  F value    Pr(>F)

X1         1 0.7770  0.7770  12.3443 0.0009618 ***

X2         1 2.5904  2.5904  41.1565 5.341e-08 ***

X3         1 6.3286  6.3286 100.5490 1.838e-13 ***

X4         1 0.0244  0.0244   0.3879 0.5362698

Residuals 49 3.0841  0.0629
```

- Full model has $P = 5$ and

$$SSE = 3.0841, \quad MSE = 0.0629, \quad R^2 = 0.7591, \quad R_a^2 = 0.7395$$

- By definition, for the full model, $C_P = P = 5$ = # of reg. coef

- Sample size $n = 54$, so for the full model:

$$AIC_P = 54 \log(3.0841/54) + 2 \times 5 = -144.5871 \text{ and}$$

$$BIC_P = 54 \log(3.0841/54) + \log(54) \times 5 = -134.6422$$

- $Press_p = 4.069$

```
> e.f=fit.f$residuals  ## residuals
> h.f=influence(fit.f)$hat  ## diagonals of hat matrix
> press.f= sum(e.f^2/(1-h.f)^2)  ## calculate press
```

Model Selection: Stepwise Regression

Model Search Procedures

(sub-models)

- The number of possible models, 2^{P-1}, grows very fast with the number potential X variables $P - 1$.
- Evaluating every possible model can be computationally infeasible even for moderate P.
- A variety of search procedures have been developed to efficiently search for the "best" model(s) in the model space.
  - *Stepwise regression procedures* greedy algorithm / strategy
  - *Best subsets algorithms*: Not applicable when the pool of potential X variables is large.

Stepwise Regression Procedures

- Use "greedy" search strategies to examine a sequence of models by adding or deleting only one X variable according to a pre-specified criterion (e.g., *AIC*) at each search step.
- Could end up with a *local optimal model* rather than the global "best" model.
- Commonly used stepwise procedures: *forward stepwise*, *forward selection*, *backward stepwise* and *backward elimination*.

Forward Stepwise Procedure

Inputs:

- A model selection criterion, e.g., *AIC*.
- An initial model $\mathbb{M}_0$, usually a small model, e.g., the null-model with no X variable. (intercept only)
- The pool of potential X variables $\mathcal{X}$. (full model), define the search space
- The set of terms that will always be in the model $\mathcal{X}_0$, e.g., the intercept term.

Starting from the initial model $\mathbb{M}_0$, at each step:

(a) Consider the X variables in the pool $\mathcal{X}$ that are not currently in the model. Examine the change of the criterion by adding each such variable into the current model.

(b) Consider the X variables that are already in the model but not in the set $\mathcal{X}_0$. Examine the change of the criterion by dropping each such variable out of the current model.

(c) Choose the operation that improves the criterion the most and update the current model accordingly.

Repeat steps (a) – (c) until there is no operation that can improve the criterion anymore.

Stopping criterion

Forward Selection and Backward Elimination

- *Forward selection* is a simplified version of forward stepwise procedure by omitting the considerations of dropping a variable currently in the model at each step.
- *Backward elimination* is the opposite of the forward selection:
  - Start with a "big" initial model, e.g., the full model.
  - At each step, examine the change of the criterion by dropping a variable currently in the model.
- *Backward stepwise procedure*: opposite of forward stepwise.

Stepwise Procedures: Comparisons

- Forward stepwise procedure often works better than forward selection when there is high multicollinearity among the potential X variables.
- Backward procedures are not good when the number of potential X variables is large. Particularly, they are not feasible when $P > n$, since then the full model can not be fitted.
- A commonly used alternative to forward stepwise procedure is to perform one pass of forward selection, followed by one pass of backward elimination.

`stepAIC()` Function in R library *MASS*

- `direction=‘‘both"` corresponds to forward stepwise procedure or backward stepwise procedure (depending on the initial model); `direction=‘‘forward"` corresponds to froward selection; `direction=‘‘backward"` corresponds to backward elimination.
- The option `scope` specifies the potential pool of X variables (`upper`) and the X variables that should always be included in the model (`lower`).
- `k=2` corresponds to *AIC* criterion; `k=log(n)` corresponds to *BIC* criterion.

Surgical Unit

```
> fit.0=lm(log(survival)~1, data=data.o) ##initial model, only intercept
> step.aic=stepAIC(fit.0, scope=list(upper=~clotting+prognostic+enzyme+liver+age+gender
+alcohol.mod+alcohol.sev, lower=~1), direction="both", k=2, trace=FALSE)
> step.aic$anova
Stepwise Model Path
Analysis of Deviance Table
Initial Model:
log(survival) ~ 1
Final Model:
log(survival) ~ enzyme + prognostic + alcohol.sev + clotting + gender + age
Step            Df   Deviance Resid. Df Resid. Dev        AIC
1                                   53  12.804509  -75.71608
2      + enzyme  1 5.47078352        52   7.333726 -103.81102
3  + prognostic  1 3.02085553        51   4.312870 -130.47855
4 + alcohol.sev  1 1.47089284        50   2.841977 -151.00214
5    + clotting  1 0.66416961        49   2.177808 -163.37593
6      + gender  1 0.09659084        48   2.081217 -163.82569
7         + age  1 0.07688125        47   2.004335 -163.85826
```

↳ intercept always be included

AIC

forward stepwise

(also consider deletion)

← null

I (X alcohol = Severe)

Model Building: Comments

For the sake of interpretability:

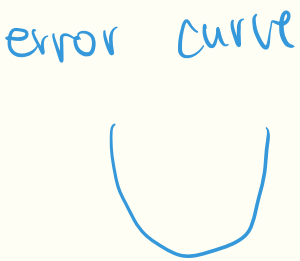
- Select all the indicator variables corresponding to a qualitative variable as a group, i.e., to be in or out of the model simultaneously.
- **Hierarchical principle**: If higher-order terms (e.g., interactions, powers) are selected, then include the related lower-order terms as well.

Model Validation

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.

Training Data vs. Validation Data

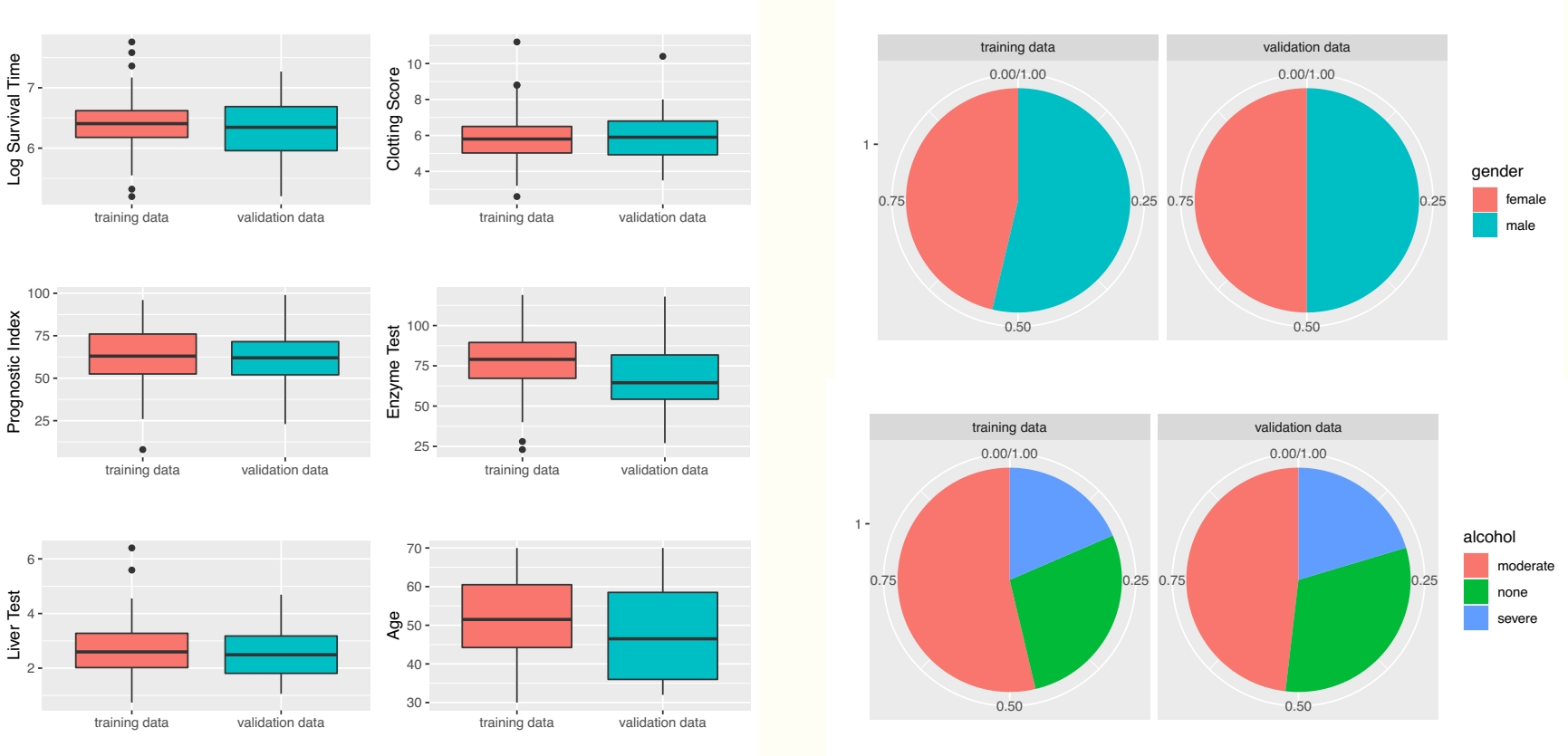


When sample size is sufficiently large, an option is to split the data into two sets, a *training data* used to build the model and a *validation data* used to check model validity.

- Training data should be sufficiently large so that a reliable model can be built from 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.

Surgical Unit: Training Data vs. Validation Data

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



Internal Validation by $Press_p$ and C_p

$Press = \sum \frac{e_i^2}{(1-h_{ii})^2} > \sum e_i^2 = SSE$

- $Press_p$ is a measure of the predictive ability of the model: $\sum (y_i - \hat{y}_{i(i)})^2$ $SSE = \sum (y_i - \hat{y}_i)^2$
  $Press_p$ not much larger than SSE_p means there is no severe over-fitting by the model. → little change . drastic change in predicting power → variance
- $C_p \approx p$ indicates little bias in the model, whereas $C_p >> p$ indicates substantial model bias. → bias

External Validation by Mean Squared Prediction Error

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

where 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 in the validation data based on the model fitted on the training data.

- $MSPE_v$ is a measure of the predictive ability of the model.
- $MSPE_v$ is usually larger than SSE/n: $MSPE_v$ not much larger than SSE/n indicates no severe over-fitting 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: Model 1 External Validation

	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

MSPE V

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

X good

	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

Change sign under this model

large

	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

largest MSPE v

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 due to relatively large SE of this coefficient.
- Models 1 and 2 perform similarly in validation.
- Based on the **principle of parsimony ("Occam's Razor")**, choose Model 1 as the final model and re-fit Model 1 on all data.

only necessary x

Surgical Unit: Model 1 Fitted on All Data

```
lm(formula = log(Y) ~ X1 + X2 + X3 + X8, data = rbind(data.o,data.v))
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 ***
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

Analysis of Variance Table
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
```

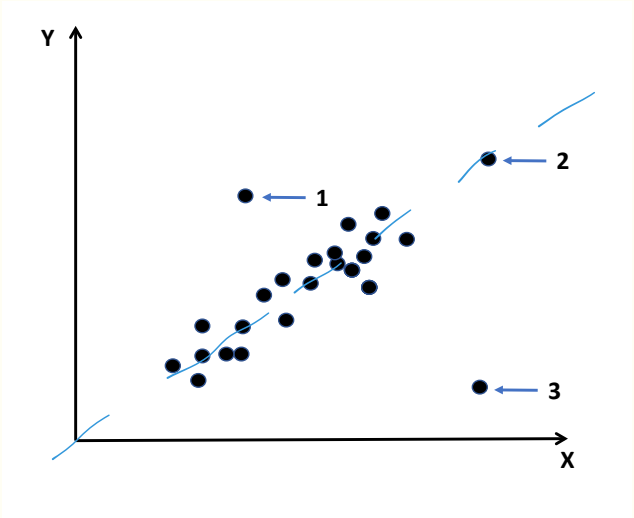
Outliers: Overview

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: outlying in Y, but not very influential since there are a few other cases with similar X values.
- Case 2: outlying in X, but not very influential since its Y value is consistent with the trend set by the majority of cases. natural extrapolation
- Cases 3: likely to be influential since it's outlying in X and its Y value is not consistent with the trend set by the majority of cases.

Identify Outlying Cases

- With one or two X variables, outlying cases can be identified by scatter plots, boxplots, etc.
- With multiple X variables, univariate outliers may not be extreme under the multivariate context; Conversely, multivariate outliers may not be detectable using single- or bivariate- analyses.
- Cases outlying in Y can be identified through residuals.
- Cases outlying in X can be identified through the diagonals of the hat matrix – *leverage* values.

Outlying in Y

Residuals

$e = (I_n - H)y$, $Var(\vec{e}) = (I_n - H) \cdot Var(y) \cdot (I_n - H)^T$

$= \sigma^2 \cdot (I_n - H)^2$

$= \sigma^2 (I_n - H)$

$\sigma^2 \cdot I_n$

assumption

- If $Var(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$, then

$$\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I}_n - \mathbf{H}), \quad \mathbf{s}^2\{\mathbf{e}\} = MSE(\mathbf{I}_n - \mathbf{H}).$$

- Variance of the *ith* residual: $\sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$ ($\in [0, 1]$, $H = (h_{ij})$)
  - Residual variances are in between 0 and σ^2.
  - The cases with larger h_{ii} have smaller residual variances.

(h_{ii}: ith diagonal of $H = X(X^TX)^{-1}X^T$)

- If the model is correct, then $\mathbf{E}\{\mathbf{e}\} = \mathbf{0}_n$.

Studentized Residuals

standardized

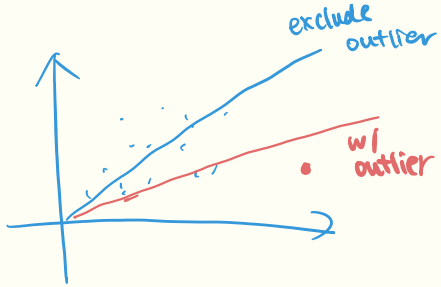
(Internally) Studentized residuals:

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

SE of e_i

- 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



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

(than e_i)

More effective in detecting outlying Y:

- The fitted regression function based on all cases could be "dragged" by the *ith* case to be close to Y_i.
- If the *ith* case is excluded, then the fitted value for the *ith* case would not be influenced by Y_i and the corresponding (deleted) residual is more likely to detect Y_i as outlying.

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

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

$\text{var}(d_i) = \text{var}\left(\frac{e_i}{1-h_{ii}}\right) = \frac{\text{var}(e_i)}{(1-h_{ii})^2}$

$= \frac{\sigma^2(1-h_{ii})}{(1-h_{ii})^2} = \frac{\sigma^2}{1-h_{ii}}$

- The larger is h_{ii}, the larger the deleted residual d_i compared with the ordinary residual e_i.

$d_i \rightarrow \frac{d_i}{\sqrt{\frac{\sigma^2}{1-h_{ii}}}}$

- 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).

we don't know σ^2, use $MSE(i)$

Studentized Deleted Residuals

a.k.a. *externally studentized residuals*:

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

SE of d_i ; $MSE_{(i)} \approx \sigma^2$

where $MSE_{(i)}$ is the MSE of the regression fit by 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

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

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

$(n-1)-p$

↳ only $n-1$ terms used

Hia: ith case is outlying in y

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

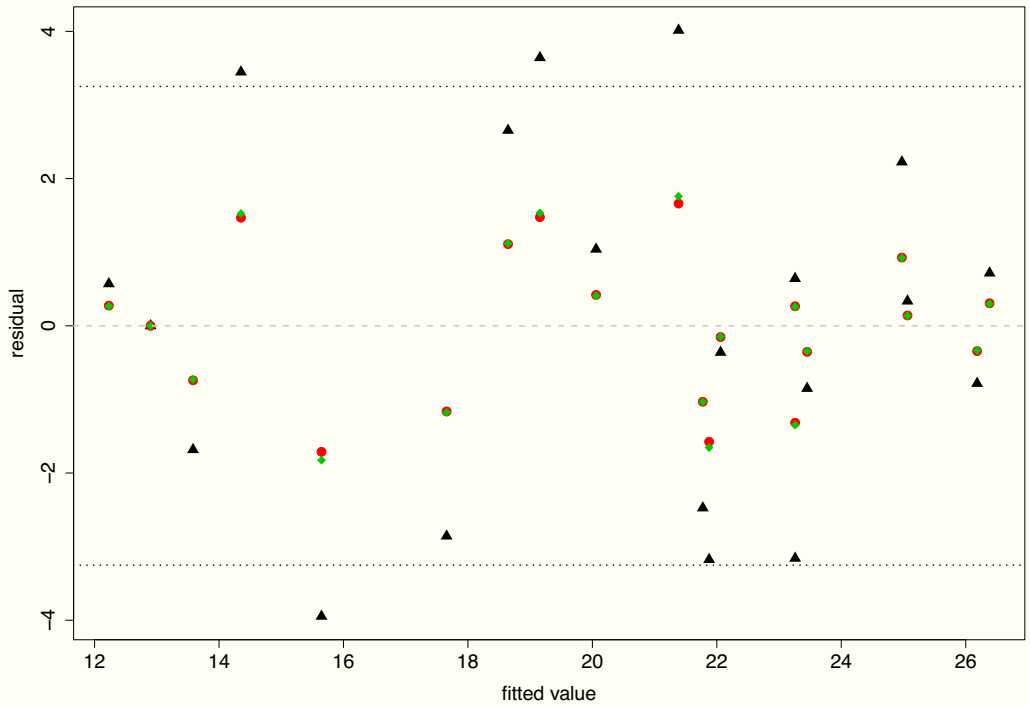
- Need to adjust for *multiple comparison*.
- Given significance level α , the **Bonferroni's procedure** controls the family-wise-type-I-error-rate at α by identifying cases with

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

replace α by α/n

as outlying Y observations.

Figure: Body Fat: Residuals vs. fitted values: Black– ordinary residuals, Red – studentized residuals, Green – studentized deleted residuals



For $\alpha = 0.1, t(1 - \alpha/40; 20 - 3 - 1) = 3.25$: No obvious outliers in Y.

Leverage and Outlying in X

Leverage Values

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

$\hat{y} = Hy$

$\hat{y}_i = H$ (ith row) $\cdot y$

$= \sum_{j=1}^{n} h_{ij} y_j$

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

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

$\sum_{j=1}^{n} h_{ij} = 1$ (because $1+1=1$)

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

Identify Outlying X by Leverage

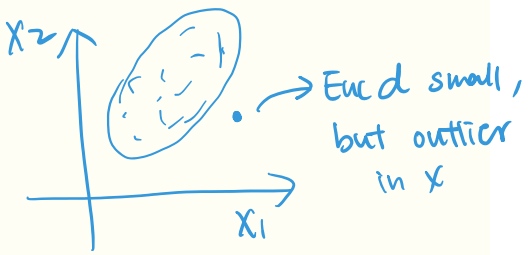
$$h_{ii} = x_i^T (X^T X)^{-1} x_i = \frac{1}{n} + \mathbf{x}_i^{*T} (\mathbf{r}_{XX})^{-1} \mathbf{x}_i^*$$

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

h_{ij} reflects the *Mahalanobis distance* between the X values of the *ith* case and the sample mean of the X values.

- A large value of h_{ij} 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 data.
- A large leverage is an indication of potential outlying in X.

Geometric Interpretation of Leverage



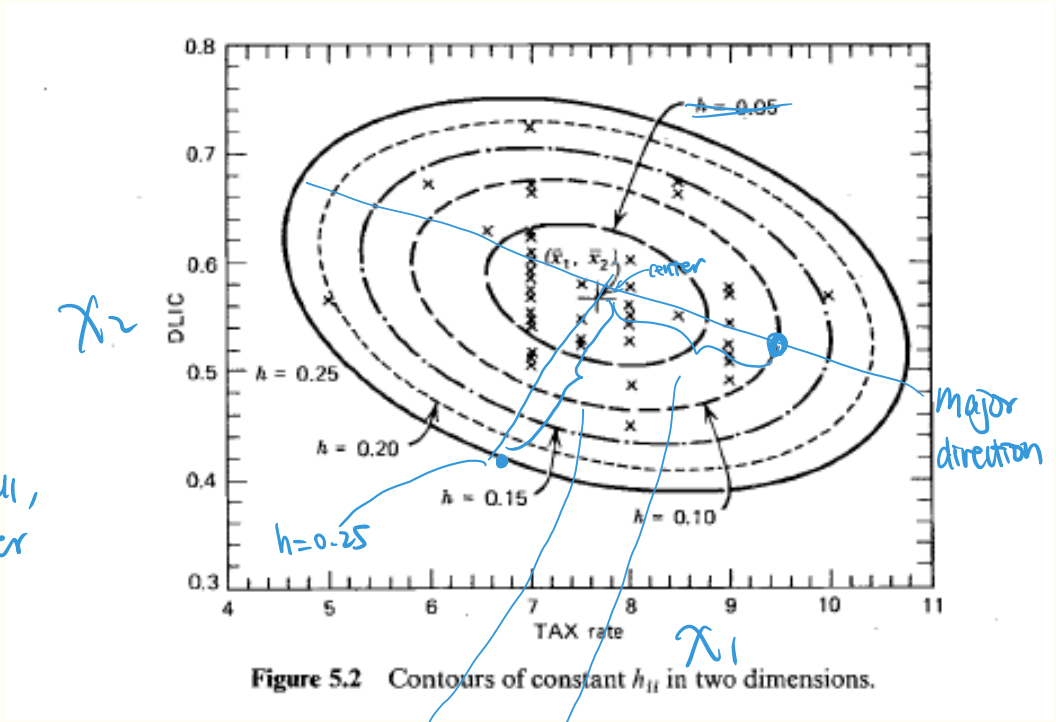


Figure 5.2 Contours of constant h_{ii} in two dimensions.

From S. Weisberg, Applied linear regression

$$r_{xx} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

$h_{ii} = x_i^T r_{xx}^{-1} x_i$

if $r_{xx} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

then $h_{ii} = \|x_i\|_2^2 + \frac{1}{n}$

if correlated, affect by h_{ii}

Having the same Euclidean distance from $\bar{\mathbf{x}}$, 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.

h_{ii} more useful than Euc. distance

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

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

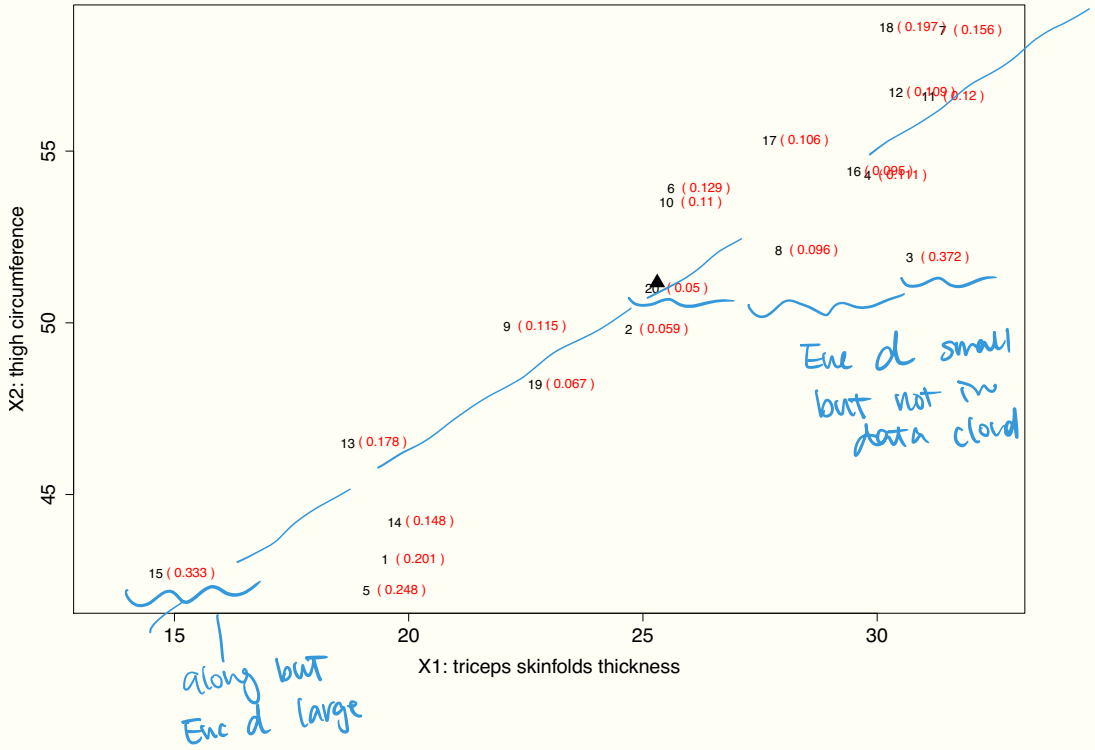
$\sum h_{ii} = \mathrm{Tr}(H) = p$

- 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 at least moderate.

Body Fat: Model 3 Leverage Values

↳ by 2 factors { distance, shape of X

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.



$\hookrightarrow \mathrm{Tr}(H) = \sum h_{ii}$

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 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: $h_{33} = 0.372$.

Hidden Extrapolation

Hidden Extrapolation

- Extrapolation occurs when predicting the response variable for X values lying outside the region of X in the data used to fit the model.
- With more than one X variables, the levels of all X variables jointly define the region of the observations.
- With two X variables, one can look at the scatter plot.
- In general, one can utilize the leverage calculation to identify extrapolation.

Identify Hidden Extrapolation by Leverage

Leverage calculation for a new X:

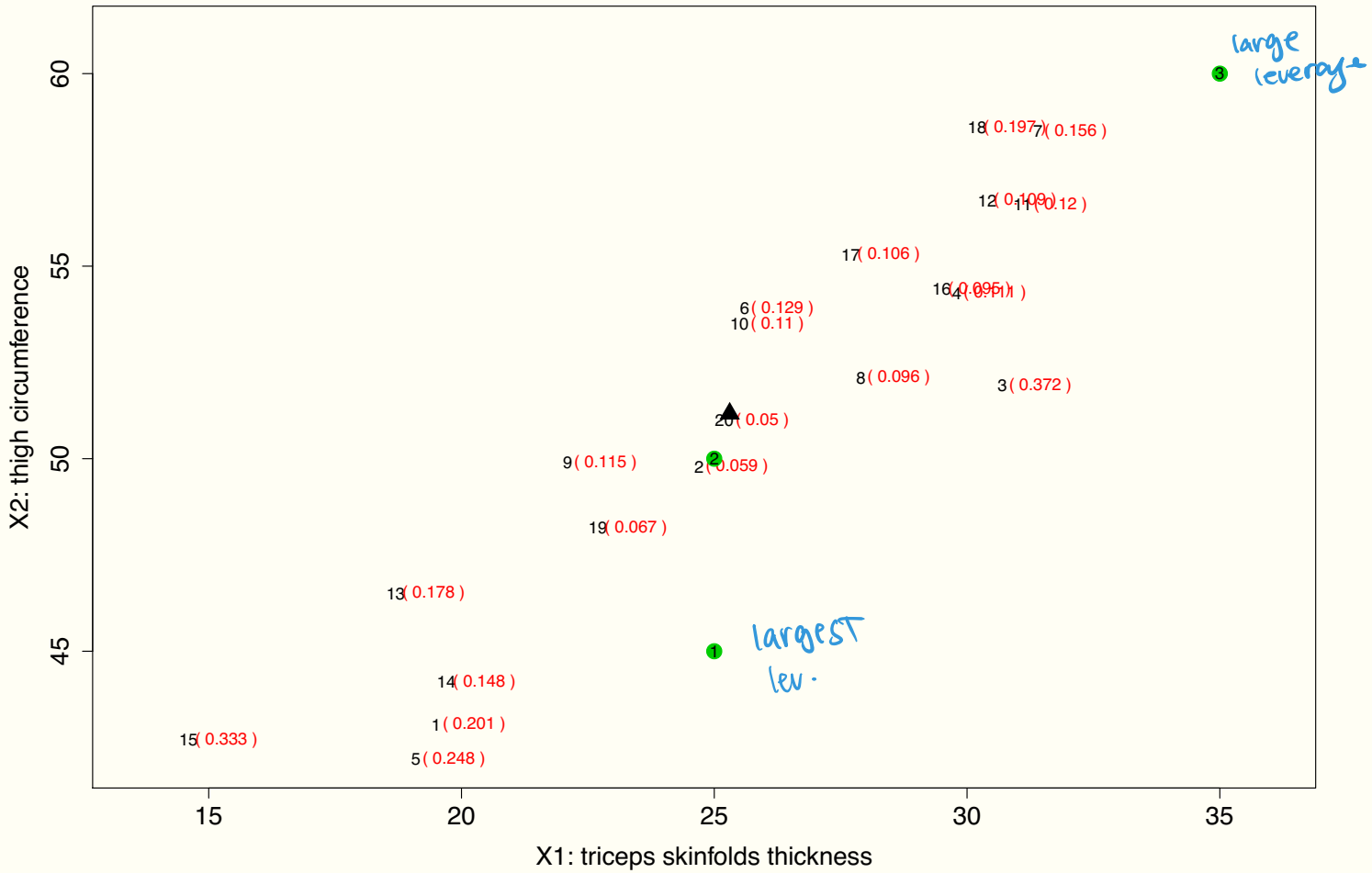
$$h_{new,new} = \mathbf{x}_{new}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}$$

- $\mathbf{x}_{new}$ is the column vector containing the new X and $\mathbf{X}$ is the design matrix of the data used to fit 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}, then extrapolation is indicated.

Body Fat: Hidden Extrapolation

```
> 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,] 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,] 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,] 0.2493753
```

Figure: Body Fat: Hidden Extrapolation



Influential Cases

Identify Influential Cases

We want to determine whether the outlying cases (in Y and/or 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:
  - 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.

Cook's Distance: Calculation

when ith case follows true model

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

$E(r_i^2) \approx 1$

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 $h_{ii} \sim p/n$).
- 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.

$\frac{h_{ii}}{1-h_{ii}} = \frac{1}{\frac{1}{h_{ii}} - 1}$ ↗ with h_{ii}

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

only define very large outlier

Body Fat: Cook's Distance

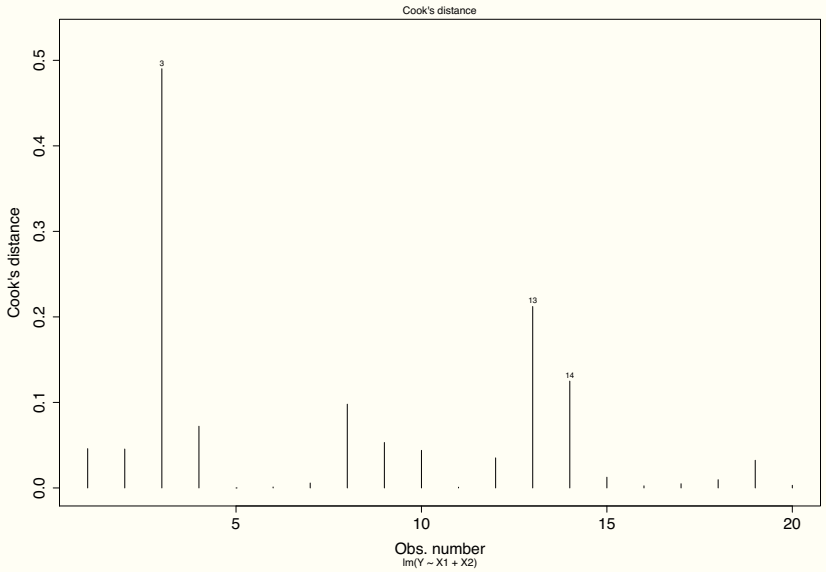
- Consider Cook's distance for case 3. It has a residual $e_3 = -3.176$ and leverage value $h_{33} = 0.372$; $p = 3$ and $MSE = 6.47$:

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

- 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.

Body Fat: Case 3 Influence

Directly examine influence of case 3:

```
> fit3=lm(Y~X1+X2, data=fat)  ## fit with all cases
> fit3.no3=lm(Y~X1+X2, data=fat[-3,]) ## fit without case 3
> per.change=abs((fit3$fitted-predict.lm(fit3.no3, fat[,1:2]))/fit3$fitted)*100
> summary(per.change) ## percentage of change in fitted values
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.6328  1.4160  2.5020  3.1060  3.9960  8.5970
```

- The main goal here is to derive a model for prediction.
- The percentage changes between fitted values based on 20 cases and those based on 19 cases (without case 3) is in between 0.63% to 8.60%. So Case 3 does not have an unduly large influence on prediction and thus may be retained.