Outline

Example: Surgical Unit

- A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 54 patients was available for analysis.
- From each patient record, the following information was extracted from the preoperation evaluation:
  - X_1 : blood clotting score
  - X_2 : prognostic index
  - X_3 : enzyme function test score
  - X_4 : liver function test score
  - X_5 : age, in years
  - X_6 : indicator variable for gender (0 = male, 1 = female)
  - X_7 : indicator variable for history of alcohol use: (0 = None, 1 = Moderate)
  - X_8 : indicator variable for history of alcohol use: (0 = None, 1 = Severe)

  These constitute the pool of potential predictor variables for a predictive regression model.
- The response variable is survival time, which was ascertained in a follow-up study.

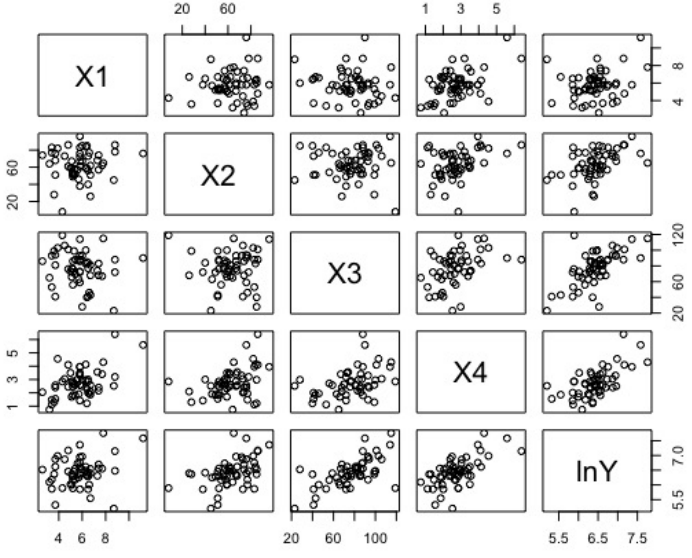
Surgical unit example

```
> mydata = read.table("SurgicalUnit.txt", header=T)
> head(mydata)
   X1 X2  X3   X4 X5 X6 X7 X8    Y   lnY
1 6.7 62  81 2.59 50  0  1  0  695 6.544
2 5.1 59  66 1.70 39  0  0  0  403 5.999
3 7.4 57  83 2.16 55  0  0  0  710 6.565
4 6.5 73  41 2.01 48  0  0  0  349 5.854
5 7.8 65 115 4.30 45  0  0  1 2343 7.759
6 5.8 38  72 1.42 65  1  1  0  348 5.852
```

Surgical unit example

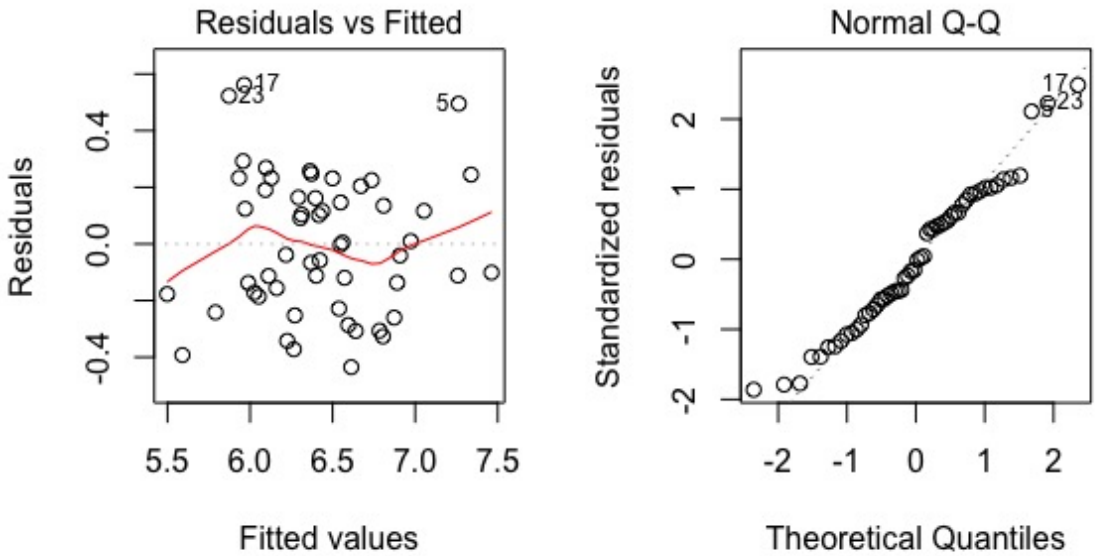
- Task 1: we consider model selection among the first 4 predictors.
- Consider log Y as response.

	X1	X2	X3	X4	lnY
X1	1.00000000	0.09011973	-0.14963411	0.5024157	0.2461879
X2	0.09011973	1.00000000	-0.02360544	0.3690256	0.4699432
X3	-0.14963411	-0.02360544	1.00000000	0.4164245	0.6538855
X4	0.50241567	0.36902563	0.41642451	1.0000000	0.6492627
lnY	0.24618787	0.46994325	0.65388548	0.6492627	1.0000000



Surgical unit example

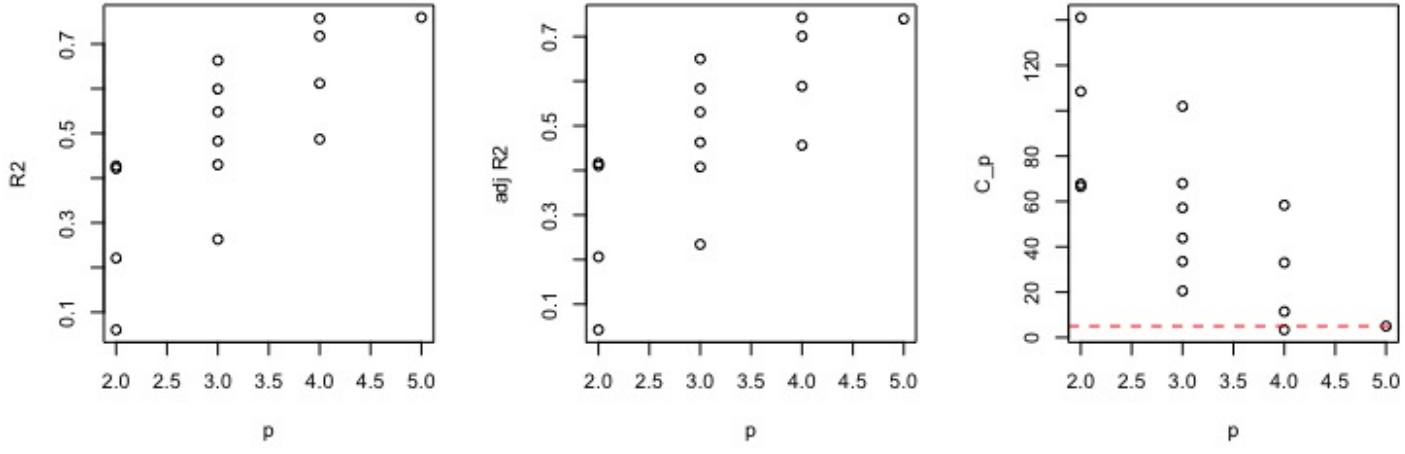
$$\ln(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon \qquad (1)$$



Surgical unit example: Model selection via best subsets

```
> source("myregsub.R")
> round(my.regsub(mydata[,1:4], lnY, nbest=6, method="exhaustive",nvmax=4),3)
  (Intercept) X1 X2 X3 X4   rsq    rss adjr2      cp     bic
1           1  0  0  1  0 0.427  7.334 0.416  66.518 -22.117
1           1  0  0  0  1 0.421  7.408 0.410  67.696 -21.574
1           1  0  1  0  0 0.221  9.974 0.206 108.469  -5.511
1           1  1  0  0  0 0.061 12.028 0.043 141.093   4.598
2           1  0  1  1  0 0.663  4.313 0.650  20.523 -46.796
2           1  0  0  1  1 0.599  5.132 0.583  33.536 -37.406
2           1  1  0  1  0 0.548  5.783 0.531  43.873 -30.961
2           1  0  1  0  1 0.483  6.620 0.463  57.175 -23.659
2           1  1  0  0  1 0.430  7.299 0.408  67.961 -18.387
2           1  1  1  0  0 0.263  9.437 0.234 101.937  -4.511
3           1  1  1  1  0 0.757  3.109 0.743   3.388 -60.489
3           1  0  1  1  1 0.718  3.615 0.701  11.434 -52.339
3           1  1  0  1  1 0.612  4.970 0.589  32.960 -35.151
3           1  1  1  0  1 0.487  6.568 0.456  58.358 -20.091
4           1  1  1  1  1 0.759  3.084 0.739   5.000 -56.926
```

Surgical unit example: Model selection via best subsets



Little increase in R^2 after three X variables are included in the model.

	(Intercept)	X1	X2	X3	X4	rsq	rss	adjr2	cp	bic
3		1	1	1	0	0.757	3.109	0.743	3.388	-60.489
3		1	0	1	1	0.718	3.615	0.701	11.434	-52.339
3		1	1	0	1	0.612	4.970	0.589	32.960	-35.151
3		1	1	1	0	0.487	6.568	0.456	58.358	-20.091

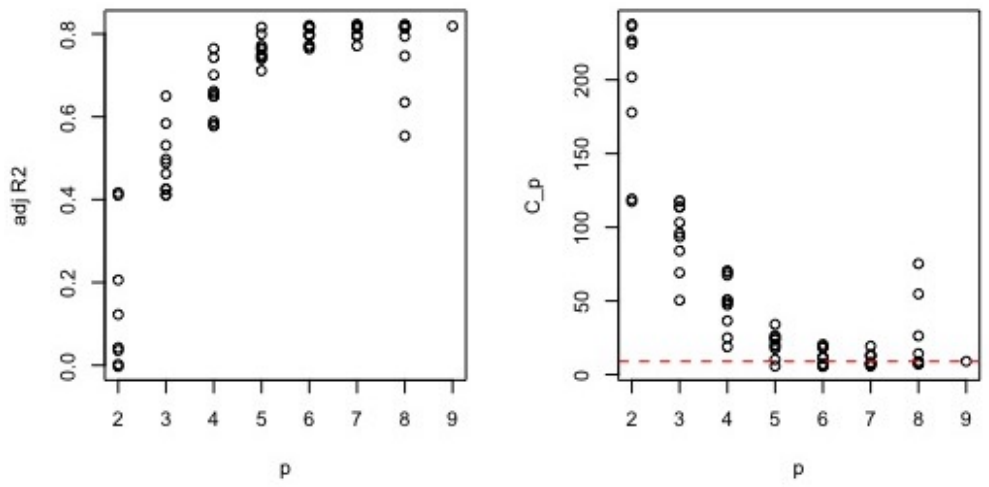
The R_4^2 criterion suggests the subset (X_1, X_2, X_3) to be reasonable.

Surgical unit example

- Task 2: perform model selection among all 8 predictors.

```
> round(my.regsub(mydata[,1:8], lnY, nbest=5, method="exhaustive",nvmax=9),3)
  (Intercept) X1 X2 X3 X4 X5 X6 X7 X8   rsq    rss adjr2      cp     bic
1           1  0  0  1  0  0  0  0  0 0.428  7.332 0.417 117.409 -22.146
1           1  0  0  0  1  0  0  0  0 0.422  7.409 0.410 119.171 -21.581
1           1  0  1  0  0  0  0  0  0 0.221  9.979 0.206 177.865  -5.498
1           1  0  0  0  0  0  0  0  1 0.139 11.028 0.122 201.811  -0.102
1           1  1  0  0  0  0  0  0  0 0.061 12.031 0.043 224.727   4.602
2           1  0  1  1  0  0  0  0  0 0.663  4.312 0.650  50.472 -46.814
2           1  0  0  1  1  0  0  0  0 0.599  5.130 0.584  69.132 -37.443
2           1  1  0  1  0  0  0  0  0 0.549  5.781 0.531  84.003 -30.989
2           1  0  0  1  0  0  0  0  1 0.516  6.194 0.497  93.434 -27.262
2           1  0  0  0  1  0  0  0  1 0.508  6.304 0.489  95.939 -26.314
3           1  0  1  1  0  0  0  0  1 0.778  2.843 0.765  18.914 -65.326
3           1  1  1  1  0  0  0  0  0 0.757  3.109 0.743  24.980 -60.502
3           1  0  1  1  1  0  0  0  0 0.718  3.614 0.701  36.525 -52.365
3           1  0  1  1  0  0  0  1  0 0.681  4.086 0.662  47.304 -45.736
3           1  0  1  1  0  1  0  0  0 0.676  4.148 0.657  48.713 -44.926
4           1  1  1  1  0  0  0  0  1 0.830  2.179 0.816   5.751 -75.704
4           1  0  1  1  1  0  0  0  1 0.814  2.377 0.799  10.267 -71.012
4           1  0  1  1  0  0  1  0  1 0.789  2.705 0.772  17.777 -64.013
4           1  0  1  1  0  1  0  0  1 0.784  2.772 0.766  19.298 -62.700
4           1  0  1  1  0  0  0  1  1 0.780  2.818 0.762  20.352 -61.808
5           1  1  1  1  0  0  1  0  1 0.837  2.082 0.821   5.541 -74.169
5           1  1  1  1  0  1  0  0  1 0.836  2.103 0.819   6.018 -73.629
5           1  1  1  1  1  0  0  0  1 0.833  2.137 0.816   6.799 -72.758
5           1  1  1  1  0  0  0  1  1 0.832  2.156 0.814   7.227 -72.287
5           1  0  1  1  1  0  1  0  1 0.818  2.333 0.799  11.261 -68.034
6           1  1  1  1  0  1  1  0  1 0.843  2.005 0.823   5.787 -72.209
6           1  1  1  1  0  0  1  1  1 0.839  2.060 0.819   7.029 -70.764
6           1  1  1  1  1  0  1  0  1 0.839  2.066 0.818   7.166 -70.607
6           1  1  1  1  0  1  0  1  1 0.838  2.069 0.818   7.246 -70.515
```

Surgical unit example



	(Intercept)	X1	X2	X3	X4	X5	X6	X7	X8	rsq	rss	adjr2	cp	bic
4	1	1	1	1	0	0	0	0	1	0.830	2.179	0.816	5.751	-75.704
5	1	1	1	1	0	0	1	0	1	0.837	2.082	0.821	5.541	-74.169
5	1	1	1	1	0	1	0	0	1	0.836	2.103	0.819	6.018	-73.629
5	1	1	1	1	1	0	0	0	1	0.833	2.137	0.816	6.799	-72.758
5	1	1	1	1	0	0	0	1	1	0.832	2.156	0.814	7.227	-72.287
6	1	1	1	1	0	1	1	0	1	0.843	2.005	0.823	5.787	-72.209
6	1	1	1	1	0	0	1	1	1	0.839	2.060	0.819	7.029	-70.764
6	1	1	1	1	1	0	1	0	1	0.839	2.066	0.818	7.166	-70.607
6	1	1	1	1	0	1	0	1	1	0.838	2.069	0.818	7.246	-70.515
6	1	1	1	1	1	1	0	0	1	0.837	2.086	0.816	7.627	-70.082
7	1	1	1	1	0	1	1	1	1	0.846	1.972	0.823	7.029	-69.121

- BIC_p: Model 1 = (X_1, X_2, X_3, X_8)
- C_p: Model 2 = $(X_1, X_2, X_3, X_6, X_8)$
- $R^2_{a,p}$: Model 3 = $(X_1, X_2, X_3, X_5, X_6, X_8)$ and Model 4 = $(X_1, X_2, X_3, X_5, X_6, X_7, X_8)$

Surgical unit example: Forward selection + AIC

```
> step(fit0, ~X1+X2+X3+X4+X5+X6+X7+X8,
direction="forward")
Start:  AIC=-75.7
lnY ~ 1

Df Sum of Sq     RSS      AIC
+ X3    1    5.4762  7.3316 -103.827
+ X4    1    5.3990  7.4087 -103.262
+ X2    1    2.8285  9.9792  -87.178
+ X8    1    1.7798 11.0279  -81.782
+ X1    1    0.7763 12.0315  -77.079
+ X6    1    0.6897 12.1180  -76.692
<none>              12.8077  -75.703
+ X5    1    0.2691 12.5386  -74.849
+ X7    1    0.2052 12.6025  -74.575
```

Step 1 X_3 enters the model.

```
Step:  AIC=-103.83
lnY ~ X3

Df Sum of Sq    RSS     AIC
+ X2    1   3.01908 4.3125 -130.48
+ X4    1   2.20187 5.1297 -121.11
+ X1    1   1.55061 5.7810 -114.66
+ X8    1   1.13756 6.1940 -110.93
<none>              7.3316 -103.83
+ X6    1   0.25854 7.0730 -103.77
+ X5    1   0.23877 7.0928 -103.61
+ X7    1   0.06498 7.2666 -102.31
```

Step 2 X_2 enters the model.

Surgical unit example: Forward selection + AIC

```
Step:  AIC=-130.48
lnY ~ X3 + X2

Df Sum of Sq    RSS     AIC
+ X8    1   1.46961 2.8429 -150.99
+ X1    1   1.20395 3.1085 -146.16
+ X4    1   0.69836 3.6141 -138.02
+ X7    1   0.22632 4.0862 -131.39
+ X5    1   0.16461 4.1479 -130.59
<none>              4.3125 -130.48
+ X6    1   0.08245 4.2300 -129.53
```

Step 3 X_8 enters the model.

```
Step:  AIC=-150.98
lnY ~ X3 + X2 + X8

Df Sum of Sq    RSS     AIC
+ X1    1   0.66408 2.1788 -163.35
+ X4    1   0.46630 2.3766 -158.66
+ X6    1   0.13741 2.7055 -151.66
<none>              2.8429 -150.99
+ X5    1   0.07081 2.7721 -150.35
+ X7    1   0.02464 2.8182 -149.46
```

Step 4 X_1 enters the model.

Surgical unit example: Forward selection + AIC

```
Step:  AIC=-163.35
lnY ~ X3 + X2 + X8 + X1

Df Sum of Sq    RSS     AIC
+ X6    1  0.096791 2.0820 -163.81
<none>             2.1788 -163.35
+ X5    1  0.075876 2.1029 -163.26
+ X4    1  0.041701 2.1371 -162.40
+ X7    1  0.022944 2.1559 -161.92
```

Step 5 X_6 enters the model.

```
Step:  AIC=-163.81
lnY ~ X3 + X2 + X8 + X1 + X6

Df Sum of Sq    RSS     AIC
+ X5    1  0.076782 2.0052 -163.83
<none>             2.0820 -163.81
+ X7    1  0.022387 2.0596 -162.39
+ X4    1  0.016399 2.0656 -162.23
```

Step 6 X_5 enters the model.

Surgical unit example: Forward selection + AIC

```
Step:  AIC=-163.83
lnY ~ X3 + X2 + X8 + X1 + X6 + X5

Df Sum of Sq    RSS     AIC
<none>              2.0052 -163.83
+ X7    1  0.033193 1.9720 -162.74
+ X4    1  0.002284 2.0029 -161.90

Call:
lm(formula = lnY ~ X3 + X2 + X8 + X1 + X6 + X5)

Coefficients:
(Intercept)           X3           X2           X8           X1           X6
4.05397      0.01512      0.01376      0.35090      0.07152      0.08732
X5
-0.00345
```

- Finally, $(X_1, X_2, X_3, X_5, X_6, X_8)$ is identified as the “best" subset of the predictor variables.

Surgical unit example: validation

- BIC_p: Model 1 = (X_1, X_2, X_3, X_8)
- C_p: Model 2 = $(X_1, X_2, X_3, X_6, X_8)$
- $R^2_{a,p}$: Model 3 = $(X_1, X_2, X_3, X_5, X_6, X_8)$ and Model 4 = $(X_1, X_2, X_3, X_5, X_6, X_7, X_8)$
- Forward regression selection: Model 3 = $(X_1, X_2, X_3, X_5, X_6, X_8)$

```
  (Intercept) X1 X2 X3 X4 X5 X6 X7 X8   rsq    rss adjr2      cp     bic
4           1  1  1  1  0  0  0  0  1 0.830  2.179 0.816   5.751 -75.704 # bic
5           1  1  1  1  0  0  1  0  1 0.837  2.082 0.821   5.541 -74.169 # cp
6           1  1  1  1  0  1  1  0  1 0.843  2.005 0.823   5.787 -72.209 # adj, automatic
7           1  1  1  1  0  1  1  1  1 0.846  1.972 0.823   7.029 -69.121 # adj
```

Based on 5-fold cross validation, all above models have similar MSPE

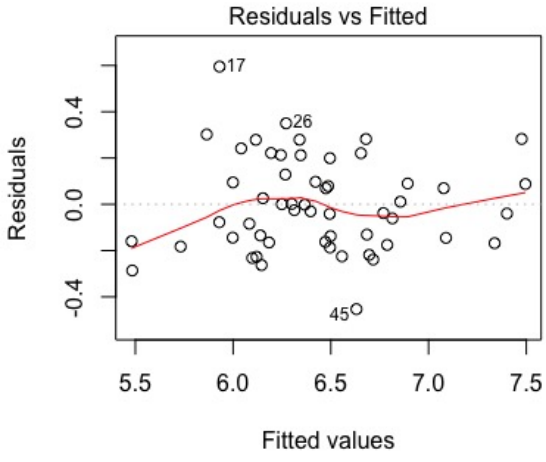
$$\text{MSPE} = \frac{1}{n^*}\sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2$$

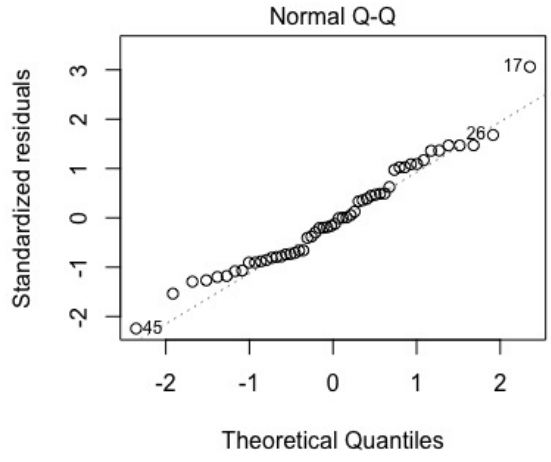
- n^* is the sample size of the validation data set.
- $\hat{Y}_i$ is the ith predicted response in the validation data set. Note that the regression coefficients in $\hat{Y}_i$ are obtained from the training data set.
- Y_i is the ith observed response in the validation data set.

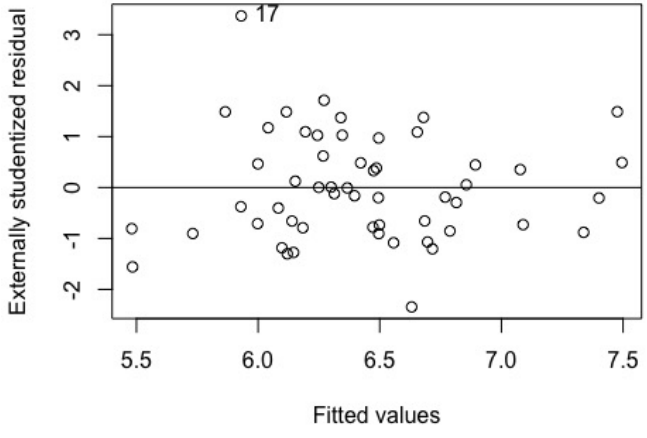
Model 1 has fewer parameters. Based on the principle of parsimony, we choose model 1 as the final model.

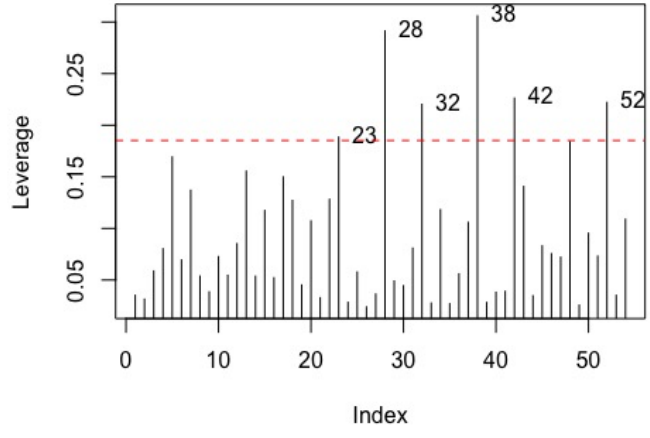
Surgical unit example: outerliers

Consider the model containing variables X_1, X_2, X_3, and X_8.



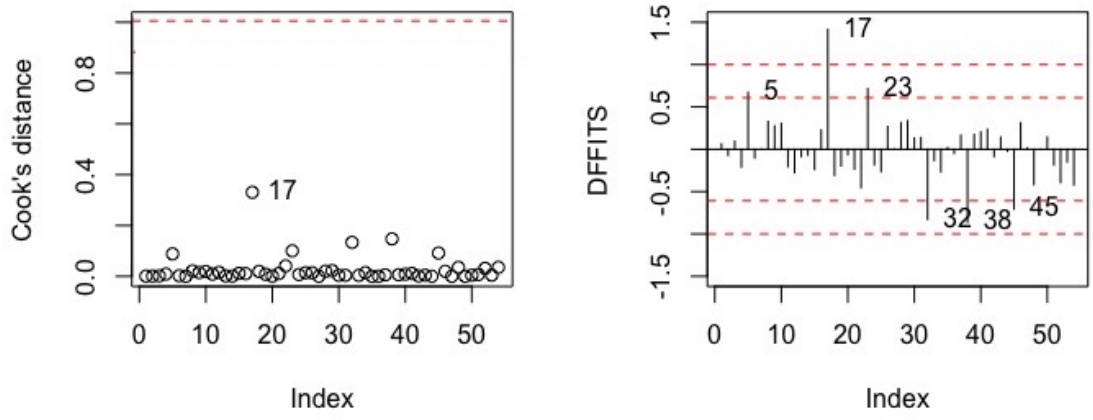






Surgical unit example: influential observations

To determine the influence of cases 17, 23, 28, 32, 38, 42, and 52, we consider their Cook's distance and DFFITS values.



Threshold: 1 or $2\sqrt{p/n} = 2\sqrt{4/54} \approx 0.54$.

The influence of the case 17 is not large enough to warrant remedial measures, and consequently the other outliers also do not appear to be overly influential.

Outline

Example: Body fat

- We consider the amount of body fat (Y) to several possible predictor variables, based on a sample of 20 healthy females.
- The possible predictor variables are
  - triceps skinfold thickness (X_1)
  - thigh circumference (X_2)
  - midarm circumference (X_3)

```
> mydata = read.table("bodyfat.txt", header=T)
> head(mydata)
    X1   X2   X3    Y
1 19.5 43.1 29.1 11.9
2 24.7 49.8 28.2 22.8
3 30.7 51.9 37.0 18.7
4 29.8 54.3 31.1 20.1
5 19.1 42.2 30.9 12.9
6 25.6 53.9 23.7 21.7
```

Body fat example: sequential SS

Source of Variation	SS	df	MS
Regression	396.98	3	132.33
X_1	352.27	1	352.27
$X_2 \mid X_1$	33.17	1	33.17
$X_3 \mid X_1, X_2$	11.54	1	11.54
Error	98.41	16	6.15
Total	495.39	19	

```
> anova(lm(Y~X1+X2+X3)). # edited output

Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X1          1 352.27  352.27 57.2768 1.131e-06 ***
X2          1  33.17   33.17  5.3931   0.03373 *
X3          1  11.55   11.55  1.8773   0.18956
Residuals 16  98.40    6.15
```

Body fat example

Q: can both X_2 and X_3 be dropped?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \tag{2}$$

- $H_0 : \beta_2 = \beta_3 = 0$
- H_a : not both β_2 and β_3 equal 0.
- The test statistic is

$$F^* = \frac{\text{SSR}(X_2, X_3|X_1)/2}{\text{SSE}(X_1, X_2, X_3)/(n-4)}$$

- Under the H_0, $F^* \sim F_{2,n-4}$.
- The decision rule is to reject H_0 at significance level α if $f^* > f_{2,n-4,\alpha}$.

Body fat example

```
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 57.2768 1.131e-06 ***
X2         1  33.17   33.17  5.3931   0.03373 *
X3         1  11.55   11.55  1.8773   0.18956
Residuals 16  98.40    6.15
```

- $H_0 : \beta_2 = \beta_3 = 0$
- $\text{SSR}(X_2, X_3|X_1) = 44.71$.
- $f^* = \frac{\text{SSR}(X_2,X_3|X_1)/2}{\text{SSE}(X_1,X_2,X_3)/(n-4)} = \frac{44.71/2}{6.15} = 3.634959$.
- P-value $\mathbb{P}(F_{2,16} > f^*) = 0.06$; at the boundary of the decision rule.
- Further analysis is required.

Body fat example

Q: Can X_3 be dropped?

```
            Df Sum Sq Mean Sq F value    Pr(>F)
X1           1 352.27  352.27 57.2768 1.131e-06 ***
X2           1  33.17   33.17  5.3931   0.03373 *
X3           1  11.55   11.55  1.8773   0.18956
Residuals 16  98.40    6.15
```

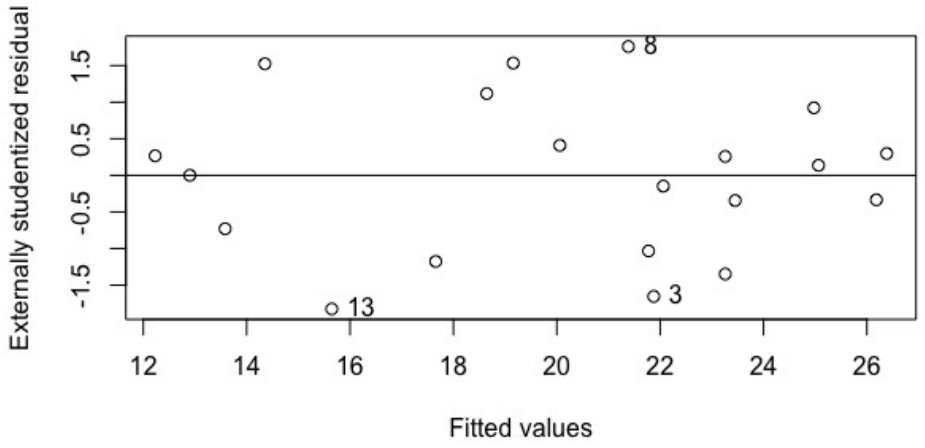
- $R^2_{Y3|12} = \frac{\text{SSR}(X_3|X_1,X_2)}{\text{SSE}(X_1,X_2)} = \frac{11.55}{109.95} = 10.5\%$. [Why?]
- F-test for $H_0 : \beta_3 = 0$ gives the p-value = 0.189.
- How does the `anova` output differ from the `lm` output?

```
> summary(lm(Y~X1+X2+X3))
                Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
X1             4.334      3.016   1.437    0.170
X2            -2.857      2.582  -1.106    0.285
X3            -2.186      1.595  -1.370    0.190
```

- Therefore we choose the two predictors (X_1, X_2) in the final model.

Body fat example: Outlier Y Observations

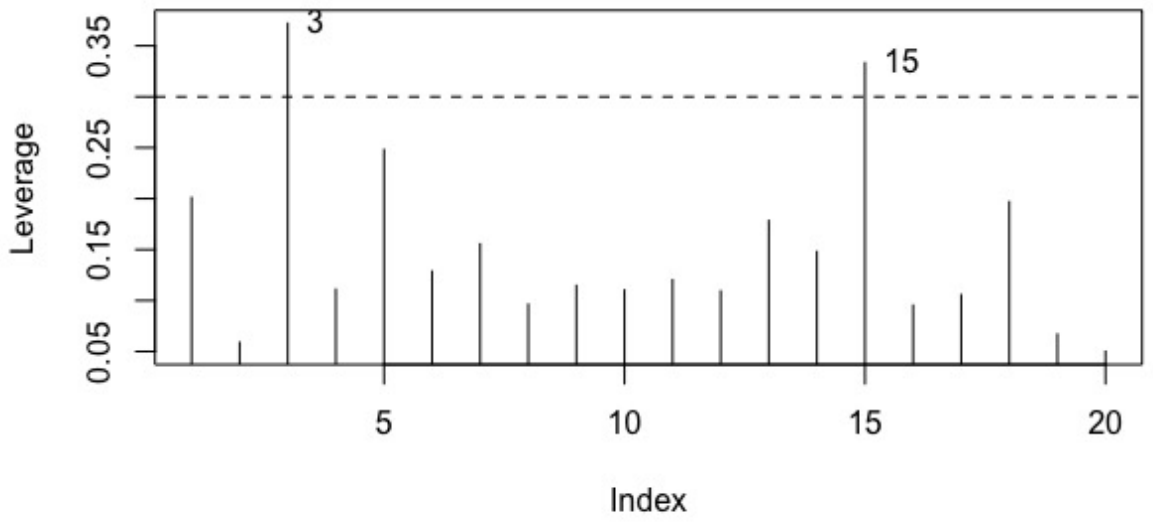
For the body fat example with two predictors (X_1, X_2), we examine whether there are outlier Y observations.



- For the Bonferroni simultaneous test procedure with a family significance level of $\alpha = 10\%$, $t_{n-p-1,1-\frac{\alpha}{2n}} = 3.252$.
- $t_3 = -1.656$, $t_8 = 1.760$, and $t_{13} = -1.825$.
- None of them are outliers.

Body fat example: Outlier in X.

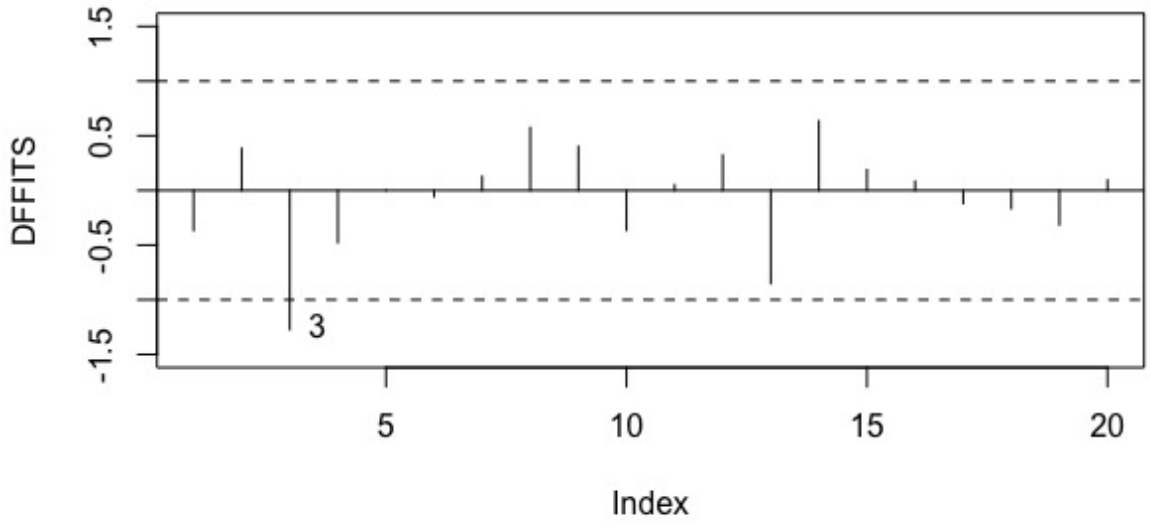
We examine whether there are outliers in X.



- $2p/n = 0.3$.
- $h_{3,3} = .372$, $h_{15,15} = .333$.

Body fat example: DFFITS

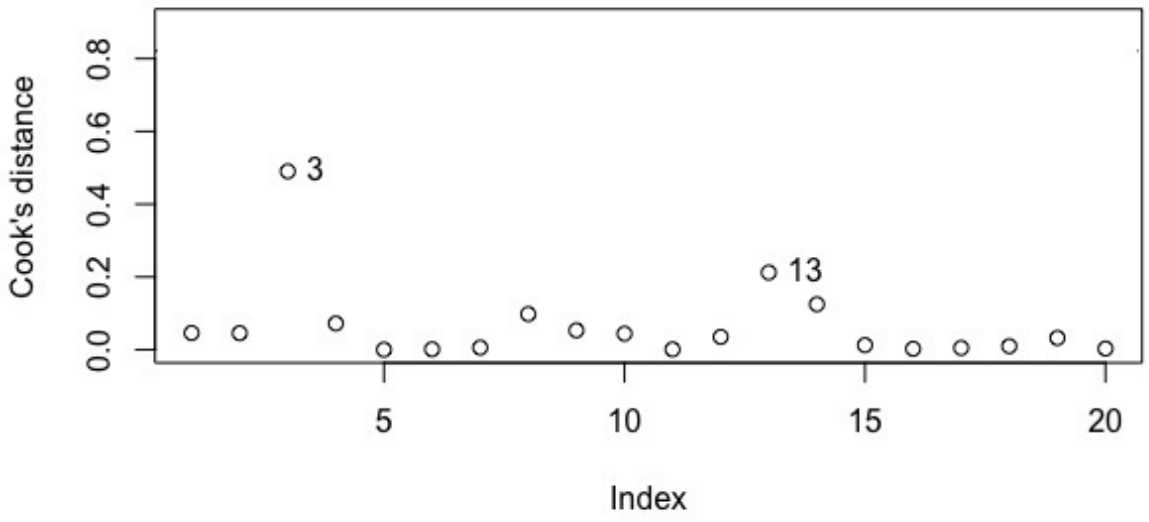
We continue with the body fat example with the predictor variables X_1 and X_2.



- $\text{DFFITS}_3 = t_3\sqrt{\frac{h_{33}}{1-h_{33}}} = -1.27,$

Though this value is larger than 1, it is close enough to 1 that the case may not be influential enough to require remedial action.

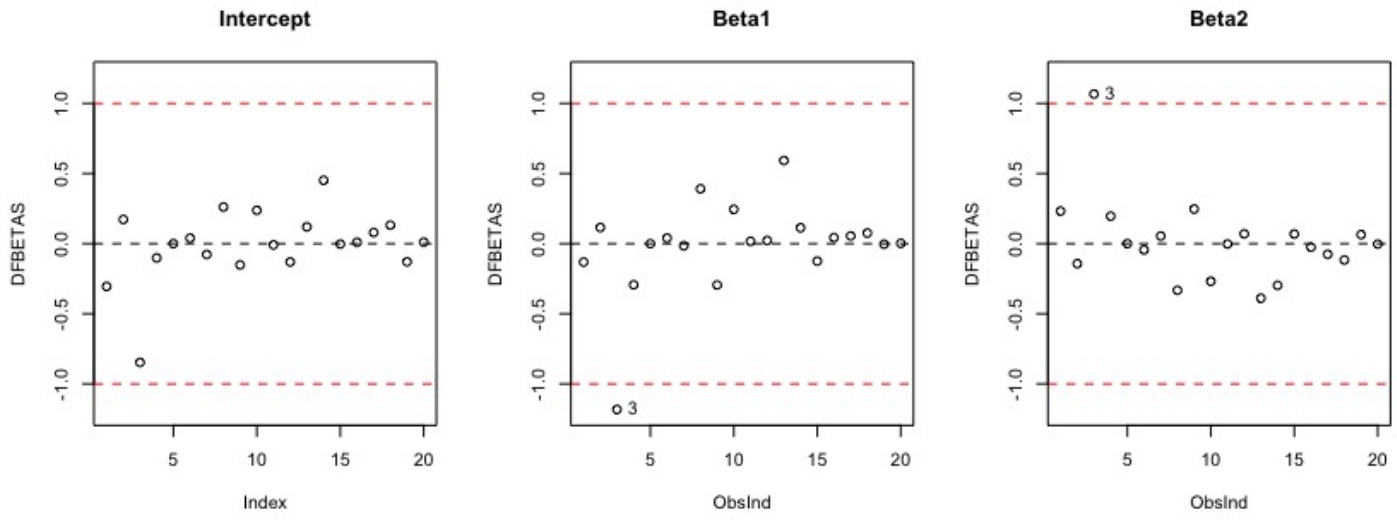
Body fat example: Cook's Distance



- Note that $e_3 = -3.176, h_{33} = .372 \Rightarrow$
  $D_3 = \frac{e_3^2}{p\hat{\sigma}^2} \frac{h_{33}}{(1-h_{33})^2} = 0.49 < 1.$

The extent of the influence for case 3 may not be large enough to call for consideration of remedial measures.

Body fat example: DFBETAS



- $\text{DFBETAS}_{1(3)} = -1.183$; $\text{DFBETAS}_{2(3)} = 1.067$
  Inclusion of case 3 leads to an increase in $\hat{\beta}_2$ but a decrease in $\hat{\beta}_1$.
- Case 3 is potentially influential; however, the DFBETAS values don't exceed 1 by very much so that case 3 may not be so influential as to require remedial action.