Diagnostics for leverage and influence

1. The hat matrix $\mathbf{H}$ is the projection of the data points onto the space spanned by $\mathbf{X}$

2. leverage(h): The diagonal elements of $\mathbf{H}$, denote by h_{ii}, $i = 1, \ldots, n$ reflect the influence of the ith data point on the model fit.

   Cutoff point for h_{ii}. If $h_i i \geq 2p/n$, then point i is judged as having an unusual value.

```
> X=matrix(c(rep(1,4), 1:4), nrow=4)
> Y=as.matrix(c(1.127,1.541,1.846, 2.407))
> beta.hat=solve(t(X) %*% X) %*% t(X) %*% Y
> H=X %*% solve(t(X) %*% X) %*% t(X)
> y.hat=H %*% Y
> X

     [,1] [,2]
[1,]    1    1
[2,]    1    2
[3,]    1    3
[4,]    1    4

> Y

      [,1]
[1,] 1.127
[2,] 1.541
[3,] 1.846
[4,] 2.407

> beta.hat

       [,1]
[1,] 0.6940
[2,] 0.4145

> H

     [,1] [,2] [,3] [,4]
[1,]  0.7  0.4  0.1 -0.2
[2,]  0.4  0.3  0.2  0.1
[3,]  0.1  0.2  0.3  0.4
[4,] -0.2  0.1  0.4  0.7

> y.hat
```

```
      [,1]
[1,] 1.108
[2,] 1.523
[3,] 1.937
[4,] 2.352
```

Some common used residuals:

standardized residuals:

studentized residuals (internally studentized residuals):

deleted residuals

Rstudentized residuals (externally studentized residuals, Jackknifed residuals):

$$d_i = \frac{e_i}{\sqrt{\hat{\sigma^2}(1-h_{ii})}}$$

$$e_{(i)} = \frac{e_i}{1-h_{ii}}$$

$$t_i = \frac{e_{(i)}}{\sqrt{\mathrm{Var}(e_{(i)})}} = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1-h_{ii})}}$$

Example:delivery time

```
> library(xtable)
> delivery[1:5,]

    TIM CAS DIS
1 16.68   7 560
2 11.50   3 220
3 12.03   3 340
4 14.88   4  80
5 13.75   6 150

> y=delivery$TIM
> x1=delivery$CAS
> x2=delivery$DIS
> f=lm(y~x1+x2)
> anova(f)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x1         1   5382    5382   506.6 < 2e-16 ***
```

```
x2         1    168     168    15.8 0.00063 ***
Residuals 22    234      11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(f)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
   Min     1Q Median     3Q    Max
-5.788 -0.663  0.436  1.157  7.420

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.34123    1.09673    2.13  0.04417 *
x1           1.61591    0.17073    9.46  3.3e-09 ***
x2           0.01438    0.00361    3.98  0.00063 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.26 on 22 degrees of freedom
Multiple R-squared: 0.96,        Adjusted R-squared: 0.956
F-statistic:  261 on 2 and 22 DF,  p-value: 4.69e-16
```

```
> y.hat=fitted(f)
> ei=residuals(f)
> di=rstandard(f)
> ti=rstudent(f)
> hii=hatvalues(f)
> deleted.r=ei/(1-hii)
> press=sum(deleted.r^2)
> press
```

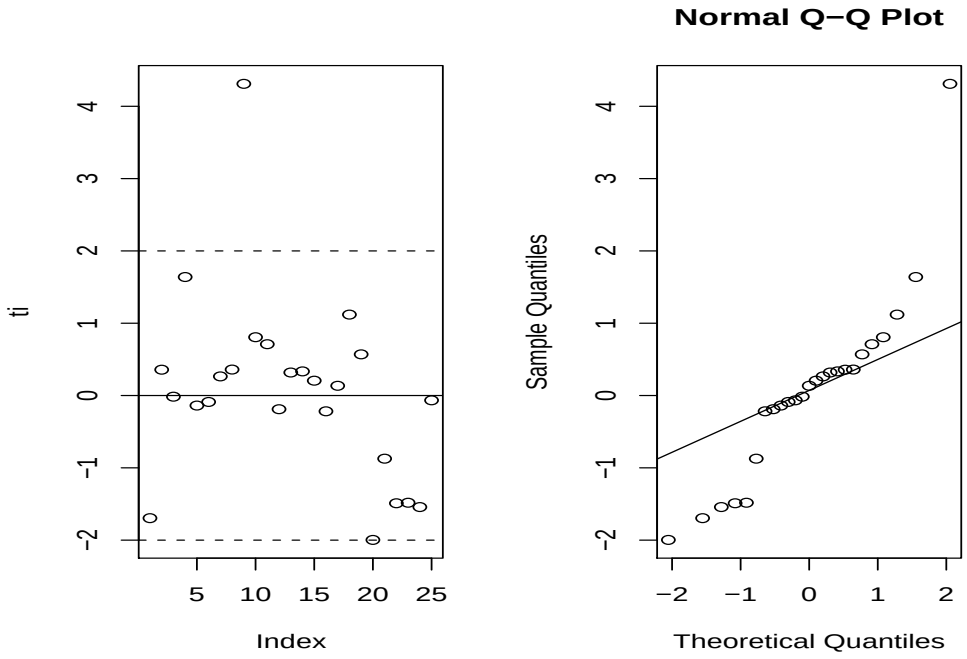
[1] 459

```
> xtable(cbind(y, y.hat, ei, di, ti, deleted.r, hii))
```

test for largest residuals

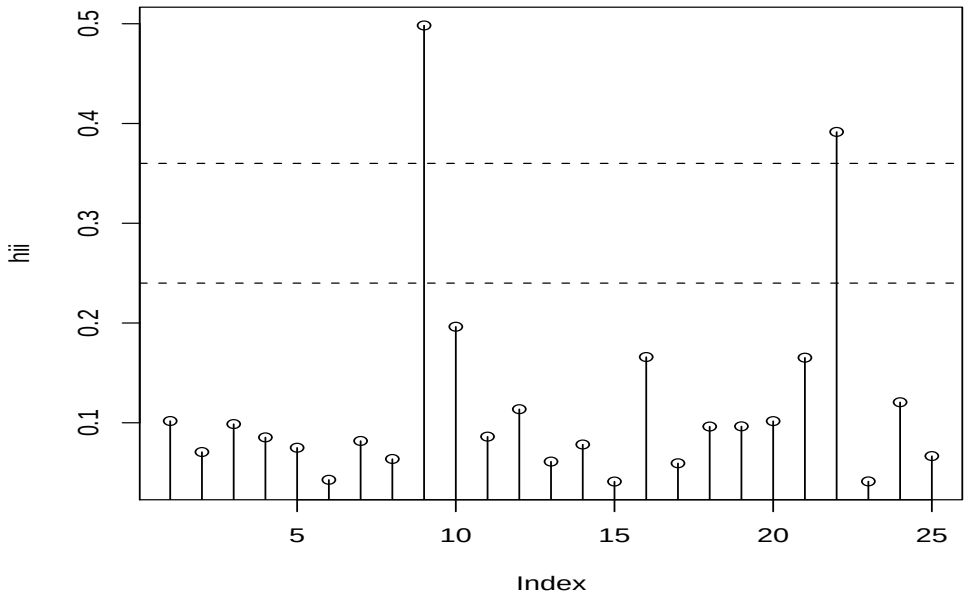
```
> outlier.test(f)

max|rstudent| = 4.311, degrees of freedom = 21,
unadjusted p = 0.000309, Bonferroni p = 0.007726

Observation: 9
```

```
> library(car)
> par(mfrow=c(1,2))
> plot(ti)
> abline(h=c(0,-2,2), lty=c(1,2,2))
> qqnorm(ti); qqline(ti) #qq.plot(f, simulate=T)  #identify few possible outliers
```

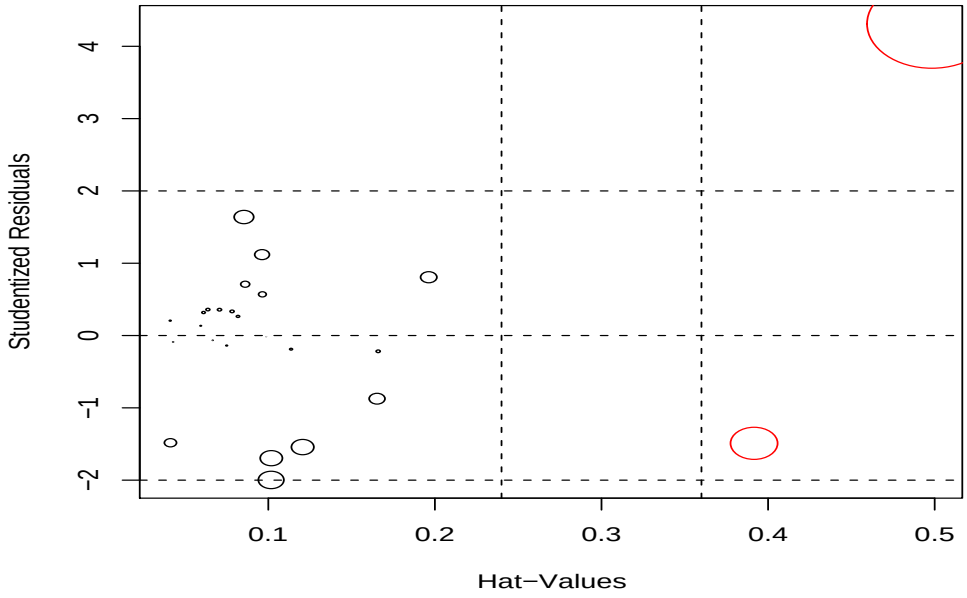


	y	y.hat	ei	di	ti	deleted.r	hii
1	16.68	21.71	-5.03	-1.63	-1.70	-5.60	0.10
2	11.50	10.35	1.15	0.36	0.36	1.23	0.07
3	12.03	12.08	-0.05	-0.02	-0.02	-0.06	0.10
4	14.88	9.96	4.92	1.58	1.64	5.38	0.09
5	13.75	14.19	-0.44	-0.14	-0.14	-0.48	0.08
6	18.11	18.40	-0.29	-0.09	-0.09	-0.30	0.04
7	8.00	7.16	0.84	0.27	0.26	0.92	0.08
8	17.83	16.67	1.16	0.37	0.36	1.24	0.06
9	79.24	71.82	7.42	3.21	4.31	14.79	0.50
10	21.50	19.12	2.38	0.81	0.81	2.96	0.20
11	40.33	38.09	2.24	0.72	0.71	2.45	0.09
12	21.00	21.59	-0.59	-0.19	-0.19	-0.67	0.11
13	13.50	12.47	1.03	0.33	0.32	1.09	0.06
14	19.75	18.68	1.07	0.34	0.33	1.16	0.08
15	24.00	23.33	0.67	0.21	0.21	0.70	0.04
16	29.00	29.66	-0.66	-0.22	-0.22	-0.79	0.17
17	15.35	14.91	0.44	0.14	0.13	0.46	0.06
18	19.00	15.55	3.45	1.11	1.12	3.82	0.10
19	9.50	7.71	1.79	0.58	0.57	1.98	0.10
20	35.10	40.89	-5.79	-1.87	-2.00	-6.44	0.10
21	17.90	20.51	-2.61	-0.88	-0.87	-3.13	0.17
22	52.32	56.01	-3.69	-1.45	-1.49	-6.06	0.39
23	18.75	23.36	-4.61	-1.44	-1.48	-4.81	0.04
24	19.83	24.40	-4.57	-1.50	-1.54	-5.20	0.12
25	10.75	10.96	-0.21	-0.07	-0.07	-0.23	0.07

```
> plot(hii, type="h")  #type for vertical line
> points(hii)
> abline(h=c(2, 3)*3/25, lty=2)
```




```
> influencePlot(f)

[1]  9 22
```



3. DEFITS residuals: The deletion influence of the ith obs on the predicted or

fitted valus

$$\frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{\hat{\sigma}^2_{(i)} h_{ii}}}$$

$\mathrm{Var}(\hat{y}) = \sigma^2 h_{ii}$ Cutoff point for DEFITS. If $\mathrm{DEFIT}_i > 2\sqrt{p/n}$, then point i is judged as having an unusual value.

```
> dffits(f)
```

```
       1        2        3        4        5        6
-0.570850  0.098619 -0.005204  0.500802 -0.039459 -0.018779
       7        8        9       10       11       12
 0.078990  0.093761  4.296081  0.398713  0.217953 -0.067670
      13       14       15       16       17       18
 0.081259  0.097363  0.042584 -0.097160  0.033916  0.365309
      19       20       21       22       23       24
 0.186168 -0.671771 -0.388501 -1.195036 -0.307539 -0.571140
      25
-0.017626
```

4. Cook's distance: The sum of squares of the deletion influences of each of the ith obs on the predicted or fitted valus

$$\frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p\hat{\sigma}^2}$$

Cutoff point for Cook distance. If $\mathrm{D}_i > qf(0.5, p, n-p)$, then point i is judged as having an unusual value.

The D_i can be rewritten as

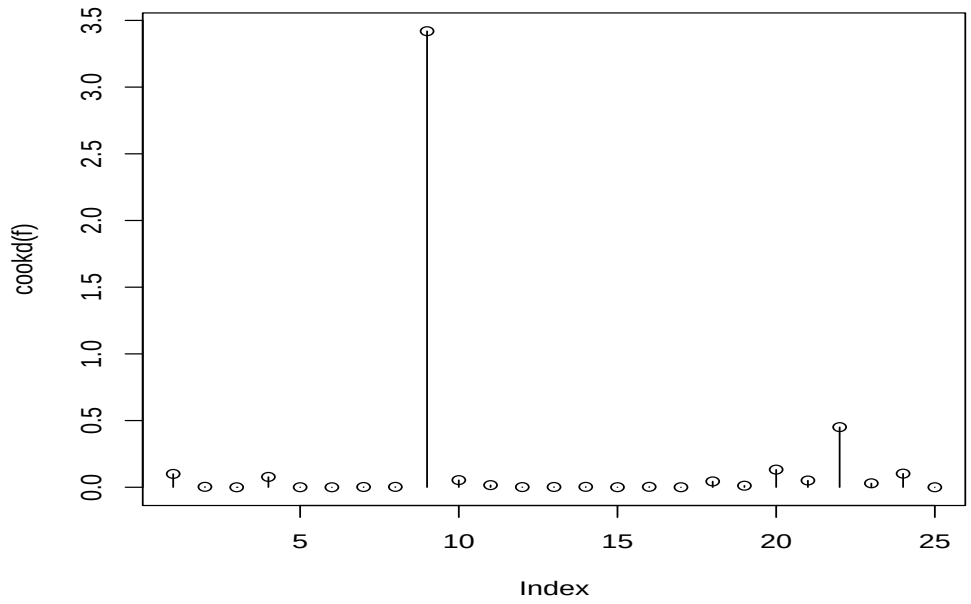
$$D_i = \frac{e_i^2}{p\hat{\sigma}^2} \frac{h_{ii}}{1 - h_{ii}}$$

Another way to interpret Cook's distance is that it is the squared Euclidean distance (apart from $p\hat{\sigma}^2$) that the vector of fitted values moves when the ith observation is deleted.

```
> cookd(f)
```

```
       1        2        3        4        5        6
1.001e-01 3.376e-03 9.456e-06 7.765e-02 5.432e-04 1.231e-04
       7        8        9       10       11       12
2.172e-03 3.051e-03 3.419e+00 5.385e-02 1.620e-02 1.596e-03
      13       14       15       16       17       18
2.295e-03 3.293e-03 6.320e-04 3.289e-03 4.013e-04 4.398e-02
      19       20       21       22       23       24
1.192e-02 1.324e-01 5.086e-02 4.510e-01 2.990e-02 1.023e-01
      25
1.085e-04
```

```
> cooks.distance(f)

        1         2         3         4         5         6
1.001e-01 3.376e-03 9.456e-06 7.765e-02 5.432e-04 1.231e-04
        7         8         9        10        11        12
2.172e-03 3.051e-03 3.419e+00 5.385e-02 1.620e-02 1.596e-03
       13        14        15        16        17        18
2.295e-03 3.293e-03 6.320e-04 3.289e-03 4.013e-04 4.398e-02
       19        20        21        22        23        24
1.192e-02 1.324e-01 5.086e-02 4.510e-01 2.990e-02 1.023e-01
       25
1.085e-04

> plot(cookd(f), type="h", pch=12)
> points(cookd(f))  #add point in the plot
```

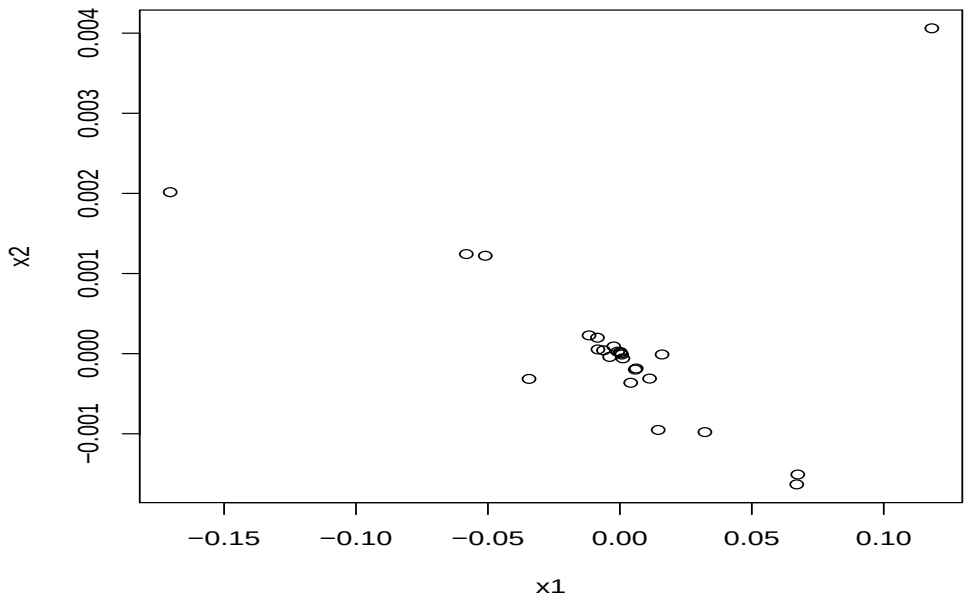


5. DFBETAS: how much the regression coefficient $\hat{\beta}_j$ changes if the, in standard deviation units, i obs were deleted.

```
> dfbs=dfbeta(f)
> head(dfbs)

  (Intercept)         x1         x2
1   -0.197151  0.0674112 -1.508e-03
2    0.100491 -0.0083216  5.314e-05
3   -0.003946  0.0006900 -1.053e-05
```

```
4    0.477707  0.0145259 -9.519e-04
5   -0.035539 -0.0023232  8.960e-05
6   -0.016477  0.0003131  3.989e-06
```

```
> plot(dfbs[,c(2,3)])
```



6. COVRATIO: ratio for the precison of estimation when ith obs were deleted.

$$\text{COVRATIO}_i = \frac{|(X'_{(i)}X_{(i)})^{-1}\hat{\sigma}^2_{(i)}|}{|(X'X)^{-1}\hat{\sigma}^2|}$$

```
> influence.measures(f)

Influence measures of
         lm(formula = y ~ x1 + x2) :

     dfb.1_   dfb.x1   dfb.x2   dffit cov.r   cook.d    hat inf
1  -0.18727  0.41131 -0.43486 -0.5709 0.871 1.00e-01 0.1018
2   0.08979 -0.04776  0.01441  0.0986 1.215 3.38e-03 0.0707
3  -0.00352  0.00395 -0.00285 -0.0052 1.276 9.46e-06 0.0987
4   0.45196  0.08828 -0.27337  0.5008 0.876 7.76e-02 0.0854
5  -0.03167 -0.01330  0.02424 -0.0395 1.240 5.43e-04 0.0750
6  -0.01468  0.00179  0.00108 -0.0188 1.200 1.23e-04 0.0429
7   0.07807 -0.02228 -0.01102  0.0790 1.240 2.17e-03 0.0818
8   0.07120  0.03338 -0.05382  0.0938 1.206 3.05e-03 0.0637
9  -2.57574  0.92874  1.50755  4.2961 0.342 3.42e+00 0.4983   *
```

```
10  0.10792 -0.33816  0.34133  0.3987 1.305 5.38e-02 0.1963
11 -0.03427  0.09253 -0.00269  0.2180 1.172 1.62e-02 0.0861
12 -0.03027 -0.04867  0.05397 -0.0677 1.291 1.60e-03 0.1137
13  0.07237 -0.03562  0.01134  0.0813 1.207 2.29e-03 0.0611
14  0.04952 -0.06709  0.06182  0.0974 1.228 3.29e-03 0.0782
15  0.02228 -0.00479  0.00684  0.0426 1.192 6.32e-04 0.0411
16 -0.00269  0.06442 -0.08419 -0.0972 1.369 3.29e-03 0.1659
17  0.02886  0.00649 -0.01570  0.0339 1.219 4.01e-04 0.0594
18  0.24856  0.18973 -0.27243  0.3653 1.069 4.40e-02 0.0963
19  0.17256  0.02357 -0.09897  0.1862 1.215 1.19e-02 0.0964
20  0.16804 -0.21500 -0.09292 -0.6718 0.760 1.32e-01 0.1017
21 -0.16193 -0.29718  0.33641 -0.3885 1.238 5.09e-02 0.1653
22  0.39857 -1.02541  0.57314 -1.1950 1.398 4.51e-01 0.3916   *
23 -0.15985  0.03729 -0.05265 -0.3075 0.890 2.99e-02 0.0413
24 -0.11972  0.40462 -0.46545 -0.5711 0.948 1.02e-01 0.1206
25 -0.01682  0.00085  0.00559 -0.0176 1.231 1.08e-04 0.0666
```

possible outliers 9 & 22

```
> cal.press=function(f){
+    ei=residuals(f)
+    hii=hatvalues(f)
+    deleted.r=ei/(1-hii)
+    press=sum(deleted.r^2)
+    (press)
+ }

> b=matrix(NA, nrow=4, ncol=6) #4*6 matrix
> head(delivery)

    TIM CAS DIS
1 16.68   7 560
2 11.50   3 220
3 12.03   3 340
4 14.88   4  80
5 13.75   6 150
6 18.11   7 330

> f=lm(TIM~CAS+DIS, data=delivery)
> summary(f)

Call:
lm(formula = TIM ~ CAS + DIS, data = delivery)

Residuals:
   Min     1Q Median     3Q    Max
```

```
-5.788 -0.663  0.436  1.157  7.420

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.34123    1.09673    2.13  0.04417 *
CAS          1.61591    0.17073    9.46  3.3e-09 ***
DIS          0.01438    0.00361    3.98  0.00063 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.26 on 22 degrees of freedom
Multiple R-squared: 0.96,        Adjusted R-squared: 0.956
F-statistic:  261 on 2 and 22 DF,  p-value: 4.69e-16
```

```
> SSres=sum(residuals(f)^2)
> p=length(coef(f))
> n=p+df.residual(f)
> SSto=(n-1)*var(delivery$TIM)
> MSres=SSres/df.residual(f)
> R2=1-(SSres/SSto)
> (sigma2=ls.diag(f)$std.dev^2)
```

```
[1] 10.62
```

```
> b[1,]=c(coef(f), MSres, R2, cal.press(f))
> del.9=delivery[-9,]
> f=lm(TIM~CAS+DIS, data=del.9)
> summary(f)
```

```
Call:
lm(formula = TIM ~ CAS + DIS, data = del.9)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0325 -1.2331  0.0199  1.4730  4.8167

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.44724    0.95247    4.67  0.00013 ***
CAS          1.49769    0.13021   11.50  1.6e-10 ***
DIS          0.01032    0.00285    3.62  0.00161 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.43 on 21 degrees of freedom
Multiple R-squared: 0.949,        Adjusted R-squared: 0.944
F-statistic:  194 on 2 and 21 DF,  p-value: 2.86e-14
```

```
> SSres=sum(residuals(f)^2)
> p=length(coef(f))
> n=p+df.residual(f)
> SSto=(n-1)*var(del.9$TIM)
> MSres=SSres/df.residual(f)
> R2=1-(SSres/SSto)
> (sigma2=ls.diag(f)$std.dev^2)

[1] 5.905

> b[2,]=c(coef(f), MSres, R2, cal.press(f))
> del.22=delivery[-22,]
> f=lm(TIM~CAS+DIS, data=del.22)
> summary(f)

Call:
lm(formula = TIM ~ CAS + DIS, data = del.22)

Residuals:
   Min     1Q Median     3Q    Max
-6.707 -0.914  0.508  1.427  5.676

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.91574    1.10511    1.73   0.0977 .
CAS          1.78632    0.20176    8.85  1.6e-08 ***
DIS          0.01237    0.00377    3.28   0.0036 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.17 on 21 degrees of freedom
Multiple R-squared: 0.956,        Adjusted R-squared: 0.952
F-statistic:  230 on 2 and 21 DF,  p-value: 5.15e-15

> SSres=sum(residuals(f)^2)
> p=length(coef(f))
> n=p+df.residual(f)
> SSto=(n-1)*var(del.22$TIM)
> MSres=SSres/df.residual(f)
> R2=1-(SSres/SSto)
> (sigma2=ls.diag(f)$std.dev^2)

[1] 10.07

> b[3,]=c(coef(f), MSres, R2, cal.press(f))
> del.both=delivery[-c(9,22),]
> f=lm(TIM~CAS+DIS, data=del.both)
> summary(f)
```

```
Call:
lm(formula = TIM ~ CAS + DIS, data = del.both)

Residuals:
   Min     1Q Median     3Q    Max
-4.060 -1.253 -0.136  1.515  5.140

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.64269    1.12598    4.12  0.00053 ***
CAS          1.45561    0.18048    8.07  1.0e-07 ***
DIS          0.01055    0.00299    3.53  0.00210 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.48 on 20 degrees of freedom
Multiple R-squared: 0.907,        Adjusted R-squared: 0.898
F-statistic: 97.8 on 2 and 20 DF,  p-value: 4.74e-11
```

```
> SSres=sum(residuals(f)^2)
> p=length(coef(f))
> n=p+df.residual(f)
> SSto=(n-1)*var(del.both$TIM)
> MSres=SSres/df.residual(f)
> R2=1-(SSres/SSto)
> (sigma2=ls.diag(f)$std.dev^2)

[1] 6.163

> b[4,]=c(coef(f), MSres, R2, cal.press(f))

> library(xtable)
> rownames(b)=c("all", "-9", "-22", "-c(9,22)")
> colnames(b)=c("b0","b1","b2","MSres", "R2", "PRESS")
> xtable(b)
```

	b0	b1	b2	MSres	R2	PRESS
all	2.34	1.62	0.01	10.62	0.96	459.04
-9	4.45	1.50	0.01	5.90	0.95	165.25
-22	1.92	1.79	0.01	10.07	0.96	474.22
-c(9,22)	4.64	1.46	0.01	6.16	0.91	186.73