Regression 10

Analysis of residuals and influence

The methods for obtaining estimates, tests and other summaries developed so far tell only half the story of regression analysis. All of these methods are computed as if the model and assumptions are correct. But, in any practical problem, some assumptions used in regression analysis are in doubt. A second phase of analysis designed to check assumptions and to build a model is usually required. In this chapter we will study methods for detecting outliers and influential observations.

1 Residuals

The residuals provide information regarding assumptions about error terms and the appropriateness of the model. Any complete data analysis requires examination of the residuals. Here we will present the outline of analysis of residuals. We will look at various residuals such as:

1. Raw residuals: $\hat{\epsilon}_i = Y_i - \hat{Y}_i$.
2. Semi-Studentized residuals: $\hat{\epsilon}_i^* = \hat{\epsilon}_i / s$.
3. PRESS residuals: $\hat{\epsilon}_{(i)} = Y_i - \hat{Y}_{(i)}$.
4. Standardized residuals: $r_i = \hat{\epsilon}_i / (s\sqrt{1 - h_{ii}})$.

(textbook: Studentized residuals, Internally Studentized residuals).

5. Studentized residuals (Jackknifed residuals): $r_{(i)} = \hat{\epsilon}_i/(s_{(i)}\sqrt{1-h_{ii}})$.

   (textbook: Studentized deleted residuals, Externally Studentized residuals).

1.1 Raw residuals

Usual regression model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, so the *true residuals*, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$, are *iid* $N(0, \sigma^2)$. But the true residuals can not be obtained because we do not know the true value of $\boldsymbol{\beta}$. We can only estimate $\boldsymbol{\epsilon}$ by

$$\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}},$$

which I call the (estimated) *raw residuals*. The raw residuals $\hat{\boldsymbol{\epsilon}} = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)'$ are consistent since $\hat{\boldsymbol{\beta}} \xrightarrow{\mathcal{P}} \boldsymbol{\beta}$ as $n \to \infty$. Thus, if n is very large relative to the number of parameters p, the raw residuals are essentially *iid* $N(0, \sigma^2)$. In small to moderate samples, however, they are not. Recall $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, which implies

$$\hat{\boldsymbol{\epsilon}} \sim N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})).$$

This tells us the followings.

1. The raw residuals $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$ are not independent because the off-diagonal elements of $(\mathbf{I} - \mathbf{H})$ are not zero.

2. The raw residuals are not identically distributed because their variances

   $$\mathrm{Var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$$

   are not equal. Here, h_{ij} is the (i,j)th element of the hat matrix $\mathbf{H}$, *i.e.*, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = [h_{ij}]$. Notice that the raw residuals have a mean of zero.

3. The variance $\mathrm{Var}(\hat{\epsilon}_i)$ decreases as x_i moves away from the center of X-range. For example, in a simple linear regression with intercept,

   $$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}. \tag{10.1}$$

   Especially when $i = j$, the h_{ii} is called the *leverage* value, or the *potential* value. As x_i moves away from $\bar{x}$, the leverage value of h_{ii} increases. Thus, $\mathrm{Var}(\hat{\epsilon}_i)$ decreases as x_i moves away from the center in X-range. The leverage is "a measure of how far from the center in X-range." This generalizes to multiple linear regression. Points with high leverage tend to decrease residuals.

4. For models with an intercept, we have $\mathbf{H1} = \mathbf{1}$, or in scalar form:

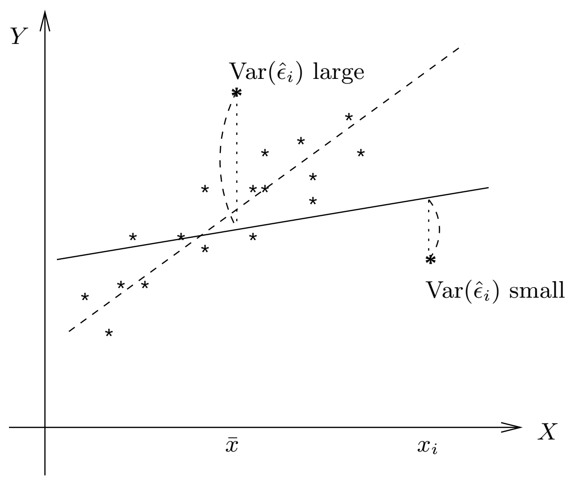
$$\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1.$$

As x_i moves far away from $\bar{x}$, the term $(x_i - \bar{x})^2/S_{xx}$ gets close to one. Thus, as can be seen from $\mathrm{Var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$, with large values of h_{ii} (*i.e.*, x_i moves far away from $\bar{x}$), $\mathrm{Var}(\hat{\epsilon}_i)$ will have a small value. We can also point this out using a scalar form of $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{HY}$:

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

In combination with $h_{ij} \approx 0$ for large n (see Eq. (10.1)) and $\sum_{j=1}^{n} h_{ij} = 1$, this shows that as the leverage h_{ii} approaches 1 (*i.e.*, $\mathbf{H} \to \mathbf{I}$), the fitted value $\hat{Y}_i$ approaches Y_i (*i.e.*, $\hat{\epsilon}_i = Y_i - \hat{Y}_i \to 0$).

Raw residuals $\hat{\epsilon}_i$ with large and small leverages



5. For very large n, all $h_{ij} \approx 0$. Thus, if $n \gg p$, then we can usually ignore the dependencies of $\hat{\epsilon}_i$.

Example 10.1. Illustration of Hat Matrix. See Table 10.2 on Page 393 of the textbook.

Minitab

Read Data

```
MTB > READ c1 c2 c11
DATA> 14 25 301
DATA> 19 32 327
DATA> 12 22 246
DATA> 11 15 187
DATA> END
4 rows read.
MTB >  name c1  'X1'
MTB >  name c2  'X2'
MTB >  name c11 'Y'
```

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

```
MTB >  REGR c11 2 c1 c2 ;
SUBC>   fits       c12 ;
SUBC>   residuals c21 ;
SUBC>   mse        k1  ;
SUBC>   hi         c24 .

Regression Analysis: Y versus X1, X2

The regression equation is
Y = 80.9 - 5.8 X1 + 11.3 X2
Predictor     Coef  SE Coef      T      P
Constant     80.93    57.94   1.40  0.396
X1           -5.84    11.74  -0.50  0.706
X2          11.325    5.931   1.91  0.307

S = 23.9768   R-Sq = 95.0%   R-Sq(adj) = 85.1%

Analysis of Variance
Source          DF       SS      MS     F      P
Regression       2  11009.9  5504.9  9.58  0.223
Residual Error   1    574.9   574.9
Total            3  11584.7

Source  DF  Seq SS
X1       1  8913.8
X2       1  2096.1

Unusual Observations
Obs    X1      Y    Fit  SE Fit  Residual  St Resid
  4  11.0  187.0  186.5    24.0       0.5      1.00 X
X denotes an observation whose X value gives it large leverage.

Residual Plots for Y
MTB > Let c33 = k1 * (1-c24)
MTB >  print c1 c2 c11 c12 c21 c24 c33

Data Display
Row  X1  X2    Y      C12      C21       C24      C33
  1  14  25  301  282.238  18.7621  0.387681  352.016
  2  19  32  327  332.292  -5.2919  0.951288   28.004
  3  12  22  246  259.951 -13.9513  0.661433  194.638
  4  11  15  187  186.519   0.4811  0.999597    0.231
```

R

Ⓡ Read Data

```
> X1 = c(14, 19, 12, 11)
> X2 = c(25, 32, 22, 15)
> Y  = c(301, 327, 246, 187)
```

Ⓡ Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

```
> source("https://raw.githubusercontent.com/AppliedStat/LM/master/Diagnostics.R")
> LM = lm (Y ~ X1 + X2)
> summary(LM)

Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   80.930     57.944   1.397    0.396
X1            -5.845     11.745  -0.498    0.706
X2            11.325      5.931   1.909    0.307

Residual standard error: 23.98 on 1 degrees of freedom
```

```
Multiple R-Squared: 0.9504, Adjusted R-squared: 0.8511
F-statistic: 9.576 on 2 and 1 DF,  p-value: 0.2228

> X = model.matrix(LM)
> X
  (Intercept) X1 X2
1           1 14 25
2           1 19 32
3           1 12 22
4           1 11 15
attr(,"assign")
[1] 0 1 2

> Y.fit = fitted(LM)

> e     = resid(LM)

> hii   = hatvalues(LM)

> mse = MSE(LM)
> mse
[1] 574.8893

> s2 = mse * (1-hii)

> cbind(X1, X2, Y, Y.fit, e, hii, s2)
  X1 X2   Y    Y.fit           e       hii          s2
1 14 25 301 282.2379  18.7620773 0.3876812 352.0155444
2 19 32 327 332.2919  -5.2918680 0.9512882  28.0038665
3 12 22 246 259.9513 -13.9512882 0.6614332 194.6384437
4 11 15 187 186.5189   0.4810789 0.9995974   0.2314369
>
> H = X %*% solve( t(X) %*% X ) %*% t(X)

> Cov.e = mse * ( diag(1, length(Y)) - H )

> Y.hat = H %*% Y

> cbind(Y.fit, Y.hat)
     Y.fit
1 282.2379 282.2379
2 332.2919 332.2919
3 259.9513 259.9513
4 186.5189 186.5189

# Note: trace(Hat matrix) = p
> sum( hii )
[1] 3
```

1.2 Internally Studentized residuals

The detection of outlying or extreme Y observations based on an examination of the residuals has been considered in earlier chapters. We used the *raw residuals* given by

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n$$

or the *semi-Studentized residuals* given by

$$\hat{\epsilon}_i^* = \frac{\hat{\epsilon}_i}{\sqrt{\text{MSE}}} = \frac{\hat{\epsilon}_i}{s},$$

where $s = \sqrt{\text{MSE}}$. The raw residuals and semi-Studentized residuals have some of the difficulty in detecting outliers when the leverages are high. The variance of $\hat{\epsilon}_i$ is $\text{Var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$ and this can be estimated by $s^2(1 - h_{ii})$. Thus, it is better to use

$$r_i = \frac{\hat{\epsilon}_i}{s\sqrt{1 - h_{ii}}}$$

which have the mean 0 and the variance 1 approximately. We call these r_i the *internally Studentized residuals*. Some textbooks, `Minitab` and `R` language call them *standardized* residuals."

Remark 10.1.

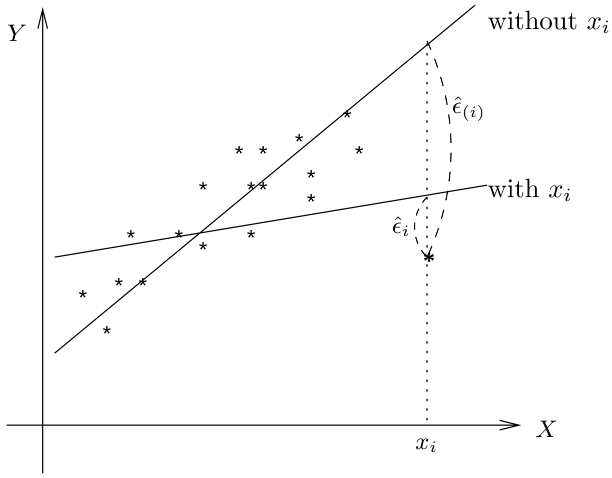
1. The internally Studentized residuals r_i are *identically distributed*, but still not independent.
2. The distribution of r_i is something like t-distribution with $\text{df} = n - p$ because σ^2 is replaced by s^2. But it is not exactly the t-distribution because the numerator and denominator are not independent. Note that we used the p normal equations to estimate the parameters $\beta_0, \ldots, \beta_{p-1}$. Hence if $n \gg p$, then we can usually ignore the dependencies of r_i.
3. For very large $n \gg p$, all $h_{ij} \approx 0$ (of course, the leverage $h_{ii} \approx 0$ also) and the r_i's are nearly proportional to $\hat{\epsilon}_i$'s. Thus, for large samples, plots of $r_1, \ldots, r_n$ look nearly the same as plots of $\hat{\epsilon}_1^*, \ldots, \hat{\epsilon}_n^*$.

△

1.3 PRESS residuals

The PRESS residual is defined as $\hat{\epsilon}_{(i)} = Y_i - \hat{Y}_{(i)}$ where $\hat{Y}_{(i)}$ is calculated with x_i omitted from the regression fit. These measure the prediction errors. Large values indicate that points are far from what the model predicts. Using the following theorem, we can calculate the PRESS residuals from the ordinary residuals.

Raw residuals $\hat{\epsilon}_i$ and PRESS residuals $\hat{\epsilon}_{(i)}$



Lemma 10.1 (Sherman-Morrison-Woodbury). *Let* $\mathbf{U}$ *and* $\mathbf{V}$ *be* $m \times k$ *matrices, and* $\mathbf{A}$ *be an* $m \times m$ *square matrix. Then we have*

$$(\mathbf{A} + \mathbf{U}\mathbf{V}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{I}_k + \mathbf{V}'\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}'\mathbf{A}^{-1},$$

where $\mathbf{I}_k$ *is an identity matrix.*

Proof. See § 2.1.3 of Golub and Van Loan (1996). □

Theorem 10.2. *The PRESS residual is obtained as*

$$\hat{\epsilon}_{(i)} = Y_i - \hat{Y}_{(i)} = \frac{\hat{\epsilon}_i}{1 - h_{ii}},$$

where h_{ii} *is the* i*th diagonal element of the hat matrix,* $\mathbf{H}$.

Proof. Let $\mathbf{x}_i'$ be the ith row of the data matrix $\mathbf{X}$ and $\mathbf{X}_{(i)}$ be the data matrix without the use of the ith observed data. Similarly, let $\mathbf{Y}_{(i)}$ be the column vector with Y_i omitted. Then we can easily show that

$$\mathbf{X}_{(i)}'\mathbf{X}_{(i)} = \mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i' \tag{10.2}$$

$$\mathbf{X}_{(i)}'\mathbf{Y}_{(i)} = \mathbf{X}'\mathbf{Y} - \mathbf{x}_iY_i. \tag{10.3}$$

It is immediate upon using Lemma 10.1 that the inverse of (10.2) is

$$\begin{aligned}
(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} &= (\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\Big(1 - \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\Big)^{-1}\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i} \\
&= (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}}. \qquad (10.4)
\end{aligned}$$

Let $\hat{\boldsymbol{\beta}}_{(i)}$ be the vector of the estimated regression coefficients with the ith observed data omitted. Then we have

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)}.$$

Then we have

$$\hat{Y}_{(i)} = \mathbf{x}'_i\hat{\boldsymbol{\beta}}_{(i)} = \mathbf{x}'_i(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)}. \qquad (10.5)$$

Substituting (10.4) into (10.5) with (10.3) gives

$$\begin{aligned}
\hat{Y}_{(i)} &= \mathbf{x}'_i\left[(\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}}\right]\mathbf{X}'_{(i)}\mathbf{Y}_{(i)} \\
&= \left[\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1} + \frac{h_{ii}\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}}\right](\mathbf{X}'\mathbf{Y} - \mathbf{x}_iY_i) \\
&= \frac{1}{1 - h_{ii}}\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y} - \mathbf{x}_iY_i) \\
&= \frac{\hat{Y}_i - h_{ii}Y_i}{1 - h_{ii}}.
\end{aligned}$$

Thus, the PRESS residual is given by

$$\hat{\epsilon}_{(i)} = Y_i - \hat{Y}_{(i)} = Y_i - \frac{\hat{Y}_i - h_{ii}Y_i}{1 - h_{ii}} = \frac{Y_i - \hat{Y}_i}{1 - h_{ii}} = \frac{\hat{\epsilon}_i}{1 - h_{ii}}.$$

□

Remark 10.2.

1. For very large $n \gg p$, all $h_{ii} \approx 0$ and $\hat{\epsilon}_{(i)} \approx \hat{\epsilon}_i$.
2. $\hat{\epsilon}_{(i)}$ far from $\hat{\epsilon}_i$ indicates influential point.

3. The PRESS residuals have unequal variances:

$$\mathrm{Var}(\hat{\epsilon}_{(i)}) = \frac{1}{(1-h_{ii})^2}\,\mathrm{Var}(\hat{\epsilon}_i) = \frac{1}{(1-h_{ii})^2}\sigma^2(1-h_{ii}) = \frac{\sigma^2}{1-h_{ii}}.$$

Dividing $\hat{\epsilon}_{(i)}$ by the estimated standard deviation $s/\sqrt{1-h_{ii}}$ gives

$$r_i = \frac{\hat{\epsilon}_i}{s\sqrt{1-h_{ii}}},$$

which is the internally Studentized residual. Thus, the standardized (internally Studentized) residuals $r_1, \ldots, r_n$ can also be thought of as standardized PRESS residuals.

△

1.4 Externally Studentized residuals

Recall that the internally Studentized residual

$$r_i = \frac{\hat{\epsilon}_i}{s\sqrt{1-h_{ii}}}, \qquad i = 1, \ldots, n$$

is not exactly t-distributed because the numerator and denominator are dependent. But if we replace $s = \sqrt{\mathrm{MSE}}$ by $s_{(i)} = \sqrt{\mathrm{MSE}_{(i)}}$, where $\mathrm{MSE}_{(i)}$ is the MSE from the model fit without the ith observation. It follows that

$$r_{(i)} = \frac{\hat{\epsilon}_i}{s_{(i)}\sqrt{1-h_{ii}}}, \qquad i = 1, \ldots, n$$

which has a t-distribution with $(n-1-p)$ degrees of freedom because $\hat{\epsilon}_i$ and $s_{(i)}$ are independent. We will call these the *externally Studentized* residuals. The textbook calls these the Studentized deleted residuals and use t_i notation instead of $r_{(i)}$. `Minitab` and R language call them "Studentized residuals."

Remark 10.3.

1. Each $r_{(i)}$ is distributed as t_{n-1-p} under the model. But $r_{(1)}, \ldots, r_{(n)}$ are *not* independent.

2. The externally Studentized residuals are traditionally used for *outlier detection* with respect to Y values.

3. Under the model, we can test the hypothesis that a single observation deviates from the model

by comparing $r_{(i)}$ to t-distribution:

$$p\text{-value} = 2 \times \text{Prob}\big[t_{n-1-p} \geq |r_{(i)}|\big]$$
$$= 2 \times \Big\{1 - \text{Prob}\big[t_{n-1-p} \leq |r_{(i)}|\big]\Big\},$$

where t_{n-1-p} is the random variable having a t-distribution with df $= n - 1 - p$. Note that even if the model holds for every observation (*i.e.*, there are no outliers), one expects about 5% of the observations to have p-values less than 0.05 when the significance level $\alpha = 5\%$ is used. So, we should not automatically call all the observations with p-values below 0.05 outliers, especially when n is large. We can conduct a formal test by means of the Bonferroni test procedure. That is, if $|r_{(i)}| > t(1 - \alpha/(2n); n - 1 - p)$, then we conclude that the ith observation is an outlier.

4. `Minitab` subcommands:

Residuals	`Subcommand`	
Raw	`RESIDUALS C21;`	$\hat{\epsilon}_i =$ `C21`
Internally Studentized	`SRESIDUALS C22;`	$r_i =$ `C22`
Externally Studentized	`TRESIDUALS C23;`	$r_{(i)} =$ `C23`
Leverage	`HI C24;`	$h_{ii} =$ `C24`

5. `R` functions:

Residuals	R Functions	Package
Raw	`resid( )`	intrinsic
Semi-Studentized	`semiresid( )`	Class Web
Internally Studentized	`rstandard( )`	intrinsic
	`stdres( )`	MASS
Externally Studentized	`rstudent( )`	intrinsic
	`studres( )`	MASS
leverage	`hatvalues( )`	intrinsic

Class Web: `https://github.com/AppliedStat/LM/blob/master/Diagnostics.R`

△

Example 10.2. Residuals, Diagonal of Hat Matrix, Studentized Deleted Residuals: Body Fat Data in Table 7.1 with Two Predictors (X_1 and X_2).

`Minitab`

Read Data

```
MTB > read c1 c2 c3 c11 ;
SUBC>      file "S:\LM\CH07TA01.txt" .
Entering data from file: S:\LM\CH07TA01.TXT
20 rows read.
MTB > name c1  'X1'
MTB > name c2  'X2'
MTB > name c11 'Y'
```

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

```
MTB > REGR c11 2 c1 c2 ;
SUBC>  residuals c21 ;
SUBC> sresiduals c22 ;
SUBC> tresiduals c23 ;
SUBC> hi          c24 .

Regression Analysis: Y versus X1, X2
The regression equation is
Y = - 19.2 + 0.222 X1 + 0.659 X2

Predictor      Coef  SE Coef       T      P
Constant    -19.174    8.361   -2.29  0.035
X1           0.2224   0.3034    0.73  0.474
X2           0.6594   0.2912    2.26  0.037

S = 2.54317   R-Sq = 77.8%   R-Sq(adj) = 75.2%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       2  385.44  192.72  29.80  0.000
Residual Error  17  109.95    6.47
Total           19  495.39

Source  DF  Seq SS
X1       1  352.27
X2       1   33.17

Residual Plots for Y

MTB > print c21 c24 c23 c22
Data Display
Row       C21       C24       C23       C22
  1  -1.68271  0.201013  -0.72999  -0.74023
  2   3.64293  0.058895   1.53425   1.47658
  3  -3.17597  0.371933  -1.65433  -1.57579
  4  -3.15847  0.110940  -1.34848  -1.31715
  5  -0.00029  0.248010  -0.00013  -0.00013
  6  -0.36082  0.128616  -0.14755  -0.15199
  7   0.71620  0.155517   0.29813   0.30645
  8   4.01473  0.096288   1.76009   1.66061
  9   2.65511  0.114636   1.11765   1.10955
 10  -2.47481  0.110244  -1.03373  -1.03165
 11   0.33581  0.120337   0.13666   0.14078
 12   2.22551  0.109266   0.92318   0.92722
 13  -3.94686  0.178382  -1.82590  -1.71215
 14   3.44746  0.148007   1.52476   1.46861
 15   0.57059  0.333212   0.26715   0.27476
 16   0.64230  0.095277   0.25813   0.26552
 17  -0.85095  0.105595  -0.34451  -0.35380
 18  -0.78292  0.196793  -0.33441  -0.34350
 19  -2.85729  0.066954  -1.17617  -1.16313
 20   1.04045  0.050085   0.40936   0.41976
```

R

Ⓡ Read Data

```
> mydata =
   read.table("https://raw.githubusercontent.com/AppliedStat/LM/master/CH07TA01.txt")
> x1 = mydata[,1]
> x2 = mydata[,2]
> y  = mydata[,4]
```

Ⓡ Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

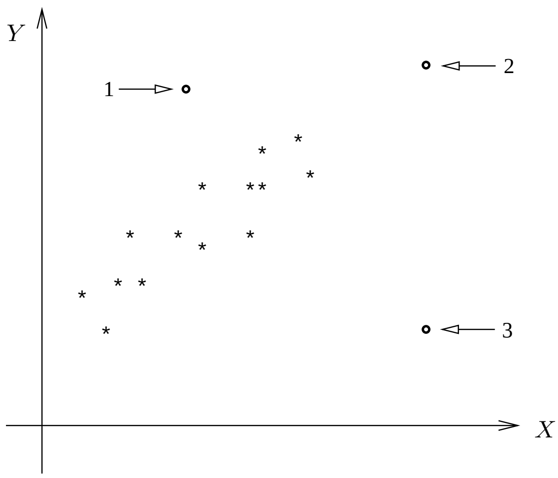
```
> source("https://raw.githubusercontent.com/AppliedStat/LM/master/Diagnostics.R")
> LM = lm ( y ~ x1 + x2 )

> e.raw = resid(LM)
> e.semi.Student = semiresid(LM)
> e.int.Student = rstandard(LM)
> e.ext.Student = rstudent(LM)
> hat.diagonal = hatvalues(LM)

> round( cbind( e.raw, hat.diagonal, e.ext.Student, e.int.Student, e.semi.Student),
   3 )
    e.raw hat.diagonal e.ext.Student e.int.Student e.semi.Student
1  -1.683        0.201        -0.730        -0.740         -0.662
2   3.643        0.059         1.534         1.477          1.432
3  -3.176        0.372        -1.654        -1.576         -1.249
4  -3.158        0.111        -1.348        -1.317         -1.242
5   0.000        0.248         0.000         0.000          0.000
6  -0.361        0.129        -0.148        -0.152         -0.142
7   0.716        0.156         0.298         0.306          0.282
8   4.015        0.096         1.760         1.661          1.579
9   2.655        0.115         1.118         1.110          1.044
10 -2.475        0.110        -1.034        -1.032         -0.973
11  0.336        0.120         0.137         0.141          0.132
12  2.226        0.109         0.923         0.927          0.875
13 -3.947        0.178        -1.826        -1.712         -1.552
14  3.447        0.148         1.525         1.469          1.356
15  0.571        0.333         0.267         0.275          0.224
16  0.642        0.095         0.258         0.266          0.253
17 -0.851        0.106        -0.345        -0.354         -0.335
18 -0.783        0.197        -0.334        -0.344         -0.308
19 -2.857        0.067        -1.176        -1.163         -1.124
20  1.040        0.050         0.409         0.420          0.409
```

‖

2 Measures of Influence

"Influence" refers to the impact of a particular observation on the model. If a single suspect observation changes our conclusions, then our conclusions are not trustworthy. We shall consider an observation to be influential if its exclusion causes major changes in the fitted regression.

Scatter plot to illustrate outlier, leverage and influence



Observation	Outlier (Y-direction)	Leverage (Outlier in X-dir)	Influential
1	$\surd$		
2		$\surd$	
3	$\surd$	$\surd$	$\surd$

2.1 Diagonals of hat matrix

The hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ has the following properties:

1. Symmetric: $\mathbf{H}' = \mathbf{H}$.
2. Idempotent: $\mathbf{HH} = \mathbf{H}$.
3. $0 \le h_{ii} \le 1$ for every i.

   With the intercept, $1/n \le h_{ii} \le 1$.
4. $\mathrm{tr}(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = \mathrm{rank}(\mathbf{X}) = p$.

In the special case of simple linear regression, we have

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$
$$\sum_{i=1}^{n} h_{ii} = \sum_{i=1}^{n} \frac{1}{n} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{S_{xx}} = 1 + 1 = 2.$$

As observations move away from the center of X-range, the leverages h_{ii} go up. Large h_{ii} indicates that an observation is *potentially* influential. The average of h_{ii} is p/n, so values of h_{ii} exceeding $2p/n$ are considered to be high leverages.

2.2 DFFITS

This statistic measures how much the fitted value for the ith observation changes when all n observations are used in fitting the regression function and when the ith observation is omitted. Denote $\hat{Y}_i$ as the fitted value for the ith observation using all n observations and $\hat{Y}_{(i)}$ as the fitted value for the ith observation with ith observation omitted. DFFITS stands for the difference between the fitted values. The DFFITS$_i$ is defined as

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s_{(i)}\sqrt{h_{ii}}},$$

where $s_{(i)} = \sqrt{\text{MSE}_{(i)}}$. Using $\hat{\epsilon}_{(i)} = Y_i - \hat{Y}_{(i)} = \frac{\hat{\epsilon}_i}{1-h_{ii}}$, we have $\hat{Y}_{(i)} = Y_i - \frac{\hat{\epsilon}_i}{1-h_{ii}}$. It follows that

$$\text{DFFITS}_i = \frac{\hat{Y}_i - Y_i + \frac{\hat{\epsilon}_i}{1-h_{ii}}}{s_{(i)}\sqrt{h_{ii}}} = \frac{\hat{\epsilon}_i \frac{h_{ii}}{1-h_{ii}}}{s_{(i)}\sqrt{h_{ii}}} = \frac{\hat{\epsilon}_i}{s_{(i)}\sqrt{1-h_{ii}}}\sqrt{\frac{h_{ii}}{1-h_{ii}}} = r_{(i)}\sqrt{\frac{h_{ii}}{1-h_{ii}}},$$

where $r_{(i)}$ is the externally Studentized residual. DFFITS$_i$ is thus a residual, inflated or shrunk by leverage.

As a guideline for identifying influential cases, we suggest considering an observation influential if the $|\text{DFFITS}_i|$ exceeds 1 for small to medium data sets and $2\sqrt{p/n}$ for large data sets (say, $n \geq 30$).

DFFITS$_i$ combines leverage h_{ii} and externally Studentized residual $r_{(i)}$ into one overall measure of how unusual an observation is.

2.3 Cook's distance

In contrast to the DFFITS_i which considers the influence of the ith observation on the fitted value $\hat{Y}_i$, Cook's distance considers the influence of the ith observation on all n fitted values. Cook's distance is defined as

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \cdot \text{MSE}},$$

where $\hat{Y}_{j(i)}$ is the fitted value for the jth observation with the ith observation omitted. Using matrix notation, it can be expressed as

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(i)})}{p \cdot \text{MSE}} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p \cdot \text{MSE}},$$

where $\hat{\mathbf{Y}}_{(i)}$ is the vector of the fitted values and $\hat{\boldsymbol{\beta}}_{(i)}$ is the vector of the estimated regression coefficients with the ith observation omitted. It has been found useful to relate D_i to the $F(p, n-p)$ distribution.

It has been suggested that observations with D_i values greater than the 50% percentile point of the F-distribution with p and $n-p$ degrees of freedom are classified as influential points. Because for most F-distributions, the 50% percentile point is near 1, the practical operational rule is to classify observations with $D_i > 1$ as being influential.

Fortunately, Cook's distance D_i can be calculated without fitting a new regression function each time a different observation is deleted. An algebraically equivalent expression is

$$D_i = \frac{\hat{\epsilon}_i^2}{p \cdot \text{MSE}} \cdot \frac{h_{ii}}{(1-h_{ii})^2} = \frac{r_i^2}{p} \cdot \frac{h_{ii}}{1-h_{ii}},$$

where r_i is the internally Studentized residual.

Cook's distance D_i combines leverage h_{ii} and internally Studentized residual r_i into one overall measure of how unusual an observation is.

2.4 DFBETAS

These statistics measure how much the values of the parameter estimates $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1})$ change when all n observations are used in estimating the regression parameters $(\beta_0, \beta_1, \ldots, \beta_{p-1})$ and when the ith observation is omitted. We also standardize these statistics by dividing them by corresponding sample standard deviations. These measures, denoted by DFBETAS, are then defined by

$$\mathrm{DFBETAS}_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{\mathrm{MSE}_{(i)} \cdot c_{kk}}},$$

where c_{kk} is the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$, $k = 0, 1, 2, \ldots, p-1$, and $i = 1, 2, \ldots, n$.

The positive/negative sign of $\mathrm{DFBETAS}_{k(i)}$ indicates that the ith observation leads to an increase/decrease in the kth parameter estimate and its absolute value indicates the amount of impact of the ith observation on the kth parameter estimate.

As a guideline for identifying influential cases, we suggest considering an observation influential if the $|\mathrm{DFBETAS}_{k(i)}|$ exceeds 1 for small to medium data sets and $2/\sqrt{n}$ for large data sets (say, $n \geq 30$).

Example 10.3. DFFITS, Cook's distances, DFBETAS – Body Fat Data with two predictors. See Table 10.4 on Page 402.

Minitab

Read Data

```
MTB > read c1 c2 c3 c11 ;
SUBC>      file "S:\LM\CH07TA01.txt" .
Entering data from file: S:\LM\CH07TA01.TXT
20 rows read.
MTB > name c1  'X1'
MTB > name c2  'X2'
MTB > name c11 'Y'
```

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

```
MTB > REGR c11 2 c1 c2 ;
SUBC>  DFITS c31 ;
SUBC>  COOKD c32 .

Regression Analysis: Y versus X1, X2
The regression equation is
Y = - 19.2 + 0.222 X1 + 0.659 X2

Predictor      Coef  SE Coef      T      P
```

```
Constant    -19.174    8.361  -2.29  0.035
X1           0.2224   0.3034   0.73  0.474
X2           0.6594   0.2912   2.26  0.037

S = 2.54317   R-Sq = 77.8%   R-Sq(adj) = 75.2%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       2  385.44  192.72  29.80  0.000
Residual Error  17  109.95    6.47
Total           19  495.39

Source  DF  Seq SS
X1       1  352.27
X2       1   33.17

Residual Plots for Y

MTB >  print c31 c32
Data Display
Row       C31       C32
  1  -0.36615  0.045951
  2   0.38381  0.045481
  3  -1.27307  0.490157
  4  -0.47635  0.072162
  5  -0.00007  0.000000
  6  -0.05669  0.001137
  7   0.12794  0.005765
  8   0.57452  0.097939
  9   0.40216  0.053134
 10  -0.36387  0.043957
 11   0.05055  0.000904
 12   0.32334  0.035154
 13  -0.85078  0.212150
 14   0.63551  0.124893
 15   0.18885  0.012575
 16   0.08377  0.002475
 17  -0.11837  0.004926
 18  -0.16553  0.009636
 19  -0.31507  0.032360
 20   0.09400  0.003097
```

R

Ⓡ Read Data

```
> mydata =
    read.table("https://raw.githubusercontent.com/AppliedStat/LM/master/CH07TA01.txt")
> x1 = mydata[,1]
> x2 = mydata[,2]
> y  = mydata[,4]
```

Ⓡ Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

```
> LM = lm ( y ~ x1 + x2 )

> DFF  = dffits (LM)
> COOK = cooks.distance (LM)
> BETA = dfbetas (LM)
>
> round ( cbind(DFF, COOK, BETA), 3)
     DFF  COOK (Intercept)     x1     x2
1  -0.366 0.046      -0.305 -0.131  0.232
2   0.384 0.045       0.173  0.115 -0.143
3  -1.273 0.490      -0.847 -1.183  1.067
4  -0.476 0.072      -0.102 -0.294  0.196
5   0.000 0.000       0.000  0.000  0.000
6  -0.057 0.001       0.040  0.040 -0.044
```

```
7    0.128 0.006       -0.078 -0.016  0.054
8    0.575 0.098        0.261  0.391 -0.332
9    0.402 0.053       -0.151 -0.295  0.247
10  -0.364 0.044        0.238  0.245 -0.269
11   0.051 0.001       -0.009  0.017 -0.002
12   0.323 0.035       -0.130  0.022  0.070
13  -0.851 0.212        0.119  0.592 -0.389
14   0.636 0.125        0.452  0.113 -0.298
15   0.189 0.013       -0.003 -0.125  0.069
16   0.084 0.002        0.009  0.043 -0.025
17  -0.118 0.005        0.080  0.055 -0.076
18  -0.166 0.010        0.132  0.075 -0.116
19  -0.315 0.032       -0.130 -0.004  0.064
20   0.094 0.003        0.010  0.002 -0.003
```

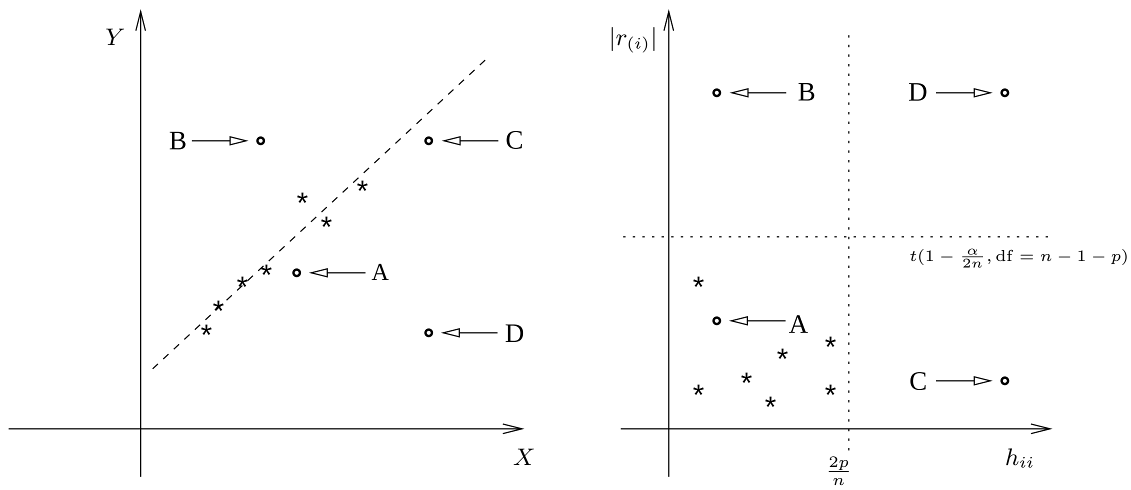
‖

2.5 Strategy to find influential observations

Influential observations can be detected by finding observations which have *high leverage* values and are *outlying* with respect to Y. Cook's distance D_i and DFFITS$_i$ combines leverage h_{ii} and outlying measure into one. They mix together deviation in X-direction with deviation in Y-direction.

My personal suggestion is to look jointly at leverages h_{ii} and externally Studentized residuals $r_{(i)}$ in a plot of $r_{(i)}$ versus h_{ii}, or a plot of $|r_{(i)}|$ versus h_{ii}.

Scatter plot and $|r_{(i)}|$ vs. h_{ii} plot



Minitab Commands

Subcommands	Residual		Note
`RESIDUALS C21;`	raw	$\hat{\epsilon}_i =$ C21	
`SRESIDUALS C22;`	internally Studentized	$r_i =$ C22	
`TRESIDUALS C23;`	externally Studentized	$r_{(i)} =$ C23	Y deviation
`HI C24;`	leverage	$h_{ii} =$ C24	X deviation
`COOKD C25;`	Cook's distance	$D_i =$ C25	X, Y mixed
`DFITS C26;`	DFFITS	$\text{DFFITS}_i =$ C26	X, Y mixed

R functions

Subcommands	Residual		Note
`resid`	raw	$\hat{\epsilon}_i$	
`semiresid`	semi-Studentized	$\hat{\epsilon}_i^*$	
`rstandard, stdres`	internally Studentized	r_i	
`rstudent, studres`	externally Studentized	$r_{(i)}$	Y deviation
`hatvalues`	leverage	h_{ii}	X deviation
`cooks.distance`	Cook's distance	D_i	X, Y mixed
`dffits`	DFFITS	DFFITS_i	X, Y mixed
`dfbetas`	DFBETAS	$\text{DFBETAS}_{k(i)}$	X, Y mixed

References

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore and London, 3rd edition.