

# Influence Diagnostics With R\*

```

> rm(list=ls())
> math =
read.table("http://www.utstat.utoronto.ca/~brunner/data/legal/math1.data.txt")
> # math = read.table("math1.data.txt") # Good for the bus
> head(math); summary(math)
  course precalc calc hsgpa hscalc hsenl ucalc frstlang sex
1 Mainstrm      2  0  78.0    65    80    39 English  F
2 Mainstrm      6  2  66.0    54    75    57 English  F
3 Mainstrm      4  4  80.2    77    70    62 English  M
4 Mainstrm      5  2  81.7    80    67    76 English  F
5 Mainstrm      4  4  86.8    87    80    86 English  M
6 Mainstrm      3  1  76.7    53    75    60 English  M
  course precalc calc hsgpa
Catch-up: 59 Min. :0.000 Min. : 0.000 Min. : 0.00
Elite : 39 1st Qu.:3.000 1st Qu.: 1.000 1st Qu.:74.50
Mainstrm:373 Median :4.000 Median : 3.000 Median :78.00
NA's :108 Mean :4.402 Mean : 3.319 Mean :75.23
3rd Qu.:6.000 3rd Qu.: 5.000 3rd Qu.:82.25
Max. :9.000 Max. :11.000 Max. :97.30
NA's :99 NA's :99 NA's :88

  hscalc hsenl ucalc frstlang sex
Min. : 0.00 Min. : 0.00 Min. : 0.0 English:402 F :266
1st Qu.: 65.00 1st Qu.:70.00 1st Qu.: 50.0 French : 5 M :285
Median : 75.00 Median :76.00 Median : 60.0 Other :144 NA's : 28
Mean : 97.32 Mean :74.44 Mean :118.6 NA's : 28
3rd Qu.: 85.00 3rd Qu.:82.00 3rd Qu.: 75.0
Max. :999.00 Max. :96.00 Max. :999.0
NA's :88 NA's :90 NA's :157
>
> # Make a new data frame without the NAs. First, how many rows and columns?
> dim(math)
[1] 579 9
>
> Math = na.omit(math); dim(Math)
[1] 321 9
> summary(Math)
  course precalc calc hsgpa
Catch-up: 14 Min. :1.000 Min. : 0.000 Min. : 0.00
Elite : 28 1st Qu.:4.000 1st Qu.: 2.000 1st Qu.:75.80
Mainstrm:279 Median :5.000 Median : 4.000 Median :79.70
Mean :4.801 Mean : 3.894 Mean :77.42
3rd Qu.:6.000 3rd Qu.: 6.000 3rd Qu.:83.80
Max. :9.000 Max. :11.000 Max. :96.20

  hscalc hsenl ucalc frstlang sex
Min. : 0.00 Min. : 0.00 Min. : 1.0 English:251 F:162
1st Qu.: 70.00 1st Qu.:71.00 1st Qu.: 50.0 French : 1 M:159
Median : 80.00 Median :77.00 Median : 61.0 Other : 69
Mean : 78.13 Mean :75.52 Mean :103.5
3rd Qu.: 86.00 3rd Qu.:83.00 3rd Qu.: 75.0
Max. :999.00 Max. :96.00 Max. :999.0
>
> # The following are impossible:
> # hscalc = 999
> # hscalc = 0
> # hsenl = 0
> # ucalc = 999
> # Make them NA and reduce the data set again.
>

```

---

\* This document is open source. See [ast page](#) for copyright information.

```

> Math$hscalcalc[Math$hscalcalc==999] = NA
> Math$hscalcalc[Math$hscalcalc==0] = NA
> Math$hsengl[Math$hsengl==0] = NA
> Math$ucalc[Math$ucalc==999] = NA
>
> Math = na.omit(Math); dim(Math)
[1] 297 9
>
> # Also make that one Francophone Other
> Math$frstlang[Math$frstlang=='French'] = 'Other'
> attach(Math)
>
> # One more problem
> contrasts(frstlang) # Built-in dummy variable coding
      French Other
English    0     0
French     1     0
Other      0     1
> frstlang = factor(frstlang); contrasts(frstlang)
      Other
English  0
Other    1
>
> # Fit a model
> fullmodell = lm(ucalc ~ course + precalc + calc + hsgpa + hscalcalc + hsengl +
frstlang + sex)
> summary(fullmodell)

```

```

Call:
lm(formula = ucalc ~ course + precalc + calc + hsgpa + hscalcalc +
    hsengl + frstlang + sex)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-92.73 -36.83 -20.23  -4.48  906.13

```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -26.8798   101.9621  -0.264  0.792
courseElite   -8.8504    53.8622  -0.164  0.870
courseMainstrm 22.2461    45.9467   0.484  0.629
precalc       10.4684     5.8985   1.775  0.077 .
calc           1.1838     3.9823   0.297  0.766
hsgpa          0.5359     0.9844   0.544  0.587
hscalcalc     -0.5950     0.9069  -0.656  0.512
hsengl         0.4007     1.1215   0.357  0.721
frstlangOther 22.2006    21.8397   1.017  0.310
sexM           7.1425    17.8935   0.399  0.690
---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 143.9 on 287 degrees of freedom
Multiple R-squared:  0.02603, Adjusted R-squared:  -0.004512
F-statistic: 0.8523 on 9 and 287 DF,  p-value: 0.5686

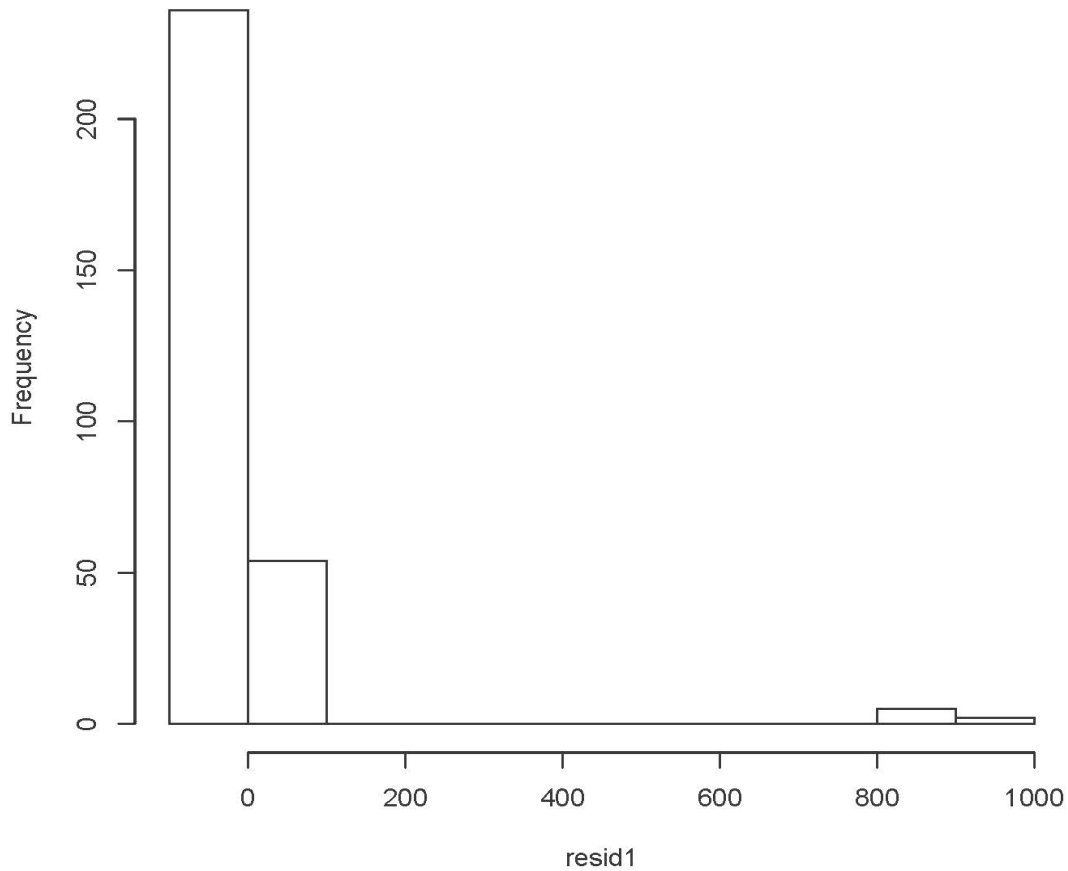
```

```

>
> # Those residuals are really strange.
> resid1 = fullmodell$residuals
> hist(resid1)

```

## Histogram of resid1



```
>
> # Track down those huge residuals
> n = length(course); n; id = 1:n
[1] 297
> crazy = id[resid1>700]; length(crazy)
[1] 7
> crazy
[1] 36 72 80 94 174 176 188
> Math[crazy,] # Just those 7 rows, all the columns
  course precalc calc hsgpa hscalc hsengl ucalc frstlang sex
63  Mainstrm      6   4  77.2    80    70   998  English  M
140 Mainstrm      4   4  81.0    65    83   998   Other  F
154 Mainstrm      7   5  85.3    89    74   998  English  M
183 Mainstrm      4   0  73.8    63    78   998   Other  M
343 Mainstrm      8   5  76.0    77    77   998  English  F
347 Mainstrm      6   8  92.7    98    84   998  English  F
365 Mainstrm      5   5  74.7    57    78   998  English  M
```

```

> sort(ucalc)
 [1] 1 2 12 13 16 17 17 18 19 19 20 22 26 26 28 28 28 30
[19] 31 31 31 32 32 33 34 34 35 35 35 36 37 37 39 39 39 39
[37] 39 39 40 41 41 42 42 43 43 44 44 45 45 46 46 46 46 46
[55] 47 47 47 47 50 50 50 50 50 50 50 50 50 50 50 50 50 50
[73] 50 51 51 51 51 51 51 51 51 51 51 52 52 52 53 53 53 53
[91] 54 54 54 54 54 54 54 55 55 55 55 55 55 55 56 56 56 56
[109] 56 57 57 57 57 57 57 57 57 57 57 58 58 58 58 58 58 60 60
[127] 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60 60
[145] 60 60 60 61 61 61 61 61 61 62 62 62 62 62 62 62 62 63 63
[163] 63 64 64 64 64 64 64 64 64 65 65 65 65 65 65 65 65 65 65
[181] 66 66 66 66 66 66 66 67 67 67 67 68 68 68 70 70 70 70 70
[199] 70 70 70 71 71 71 71 71 71 71 71 71 71 71 72 72 72 72 73
[217] 73 73 73 74 74 74 74 74 74 75 75 75 76 76 76 76 76 76 76
[235] 76 77 77 77 77 77 77 78 78 78 78 78 80 80 80 80 80 80 80
[253] 81 81 81 82 82 82 83 84 84 84 84 84 85 86 86 86 86 87 87
[271] 87 88 88 88 88 90 90 90 91 91 93 94 94 94 95 95 95 95 96
[289] 97 99 998 998 998 998 998 998 998 998
>
> # Fix it
> Math$ucalc[Math$ucalc==998] = NA
> # Fix dummy variables of frstlang at the data frame level too.
> Math$frstlang = factor(Math$frstlang)
> Math = na.omit(Math)
> dim(Math)
[1] 290 9
> attach(Math)
The following object is masked _by_ .GlobalEnv:

    frstlang

The following objects are masked from Math (pos = 3):

    calc, course, frstlang, hscal, hsengl, hsgpa, precalc, sex,
    ucalc

>
> # Fit the model again, still calling it fullmodell
> fullmodell = lm(ucalc ~ course + precalc + calc + hsgpa + hscal + hsengl +
frstlang + sex)
Error in model.frame.default(formula = ucalc ~ course + precalc + calc + :
variable lengths differ (found for 'frstlang')
>
> length(frstlang)
[1] 297
> rm(frstlang)
> length(frstlang)
[1] 290

```

```
>
> fullmodell = lm(ucalc ~ course + precalc + calc + hsgpa + hscal + hsengl +
frstlang + sex)
> summary(fullmodell)
```

```
Call:
lm(formula = ucalc ~ course + precalc + calc + hsgpa + hscal +
    hsengl + frstlang + sex)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-51.801  -6.911   1.825   8.845  31.892
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -26.62766   10.25376  -2.597  0.009905 **
courseElite   -8.92387    5.40848  -1.650  0.100069
courseMainstrm -5.78575    4.61603  -1.253  0.211104
precalc       2.03123    0.59845   3.394  0.000788 ***
calc          0.77213    0.40292   1.916  0.056337 .
hsgpa         0.31636    0.09886   3.200  0.001532 **
hscal         0.58927    0.09270   6.357  8.33e-10 ***
hsengl        0.10509    0.11304   0.930  0.353356
frstlangOther 5.81009    2.24539   2.588  0.010170 *
sexM          -1.35617    1.82179  -0.744  0.457250
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 14.43 on 280 degrees of freedom
Multiple R-squared:  0.3996, Adjusted R-squared:  0.3803
F-statistic: 20.71 on 9 and 280 DF, p-value: < 2.2e-16
```

```
>
> # That is a relief! At this point it is messy enough to
> # justify a clean re-start. Quit R.
```

```

> # Read and fix the data
> rm(list=ls())
> math =
read.table("http://www.utstat.utoronto.ca/~brunner/data/legal/math1.data.txt")
> # Clean the data
> math$hscalcl[math$hscalcl==999] = NA
> math$hscalcl[math$hscalcl==0] = NA
> math$hsengl[math$hsengl==0] = NA
> math$ucalcl[math$ucalcl==999] = NA
> math$ucalcl[math$ucalcl==998] = NA
> math$frstlang[math$frstlang=='French'] = 'Other'
> math$frstlang = factor(math$frstlang)
> math = na.omit(math)
> attach(math)
>
> fullmodell = lm(ucalcl ~ course + precalcl + calcl + hsgpa + hscalcl + hsengl +
frstlang + sex)
> summary(fullmodell)

```

Call:

```
lm(formula = ucalcl ~ course + precalcl + calcl + hsgpa + hscalcl +
hsengl + frstlang + sex)
```

Residuals:

Min	1Q	Median	3Q	Max
-51.801	-6.911	1.825	8.845	31.892

Coefficients:

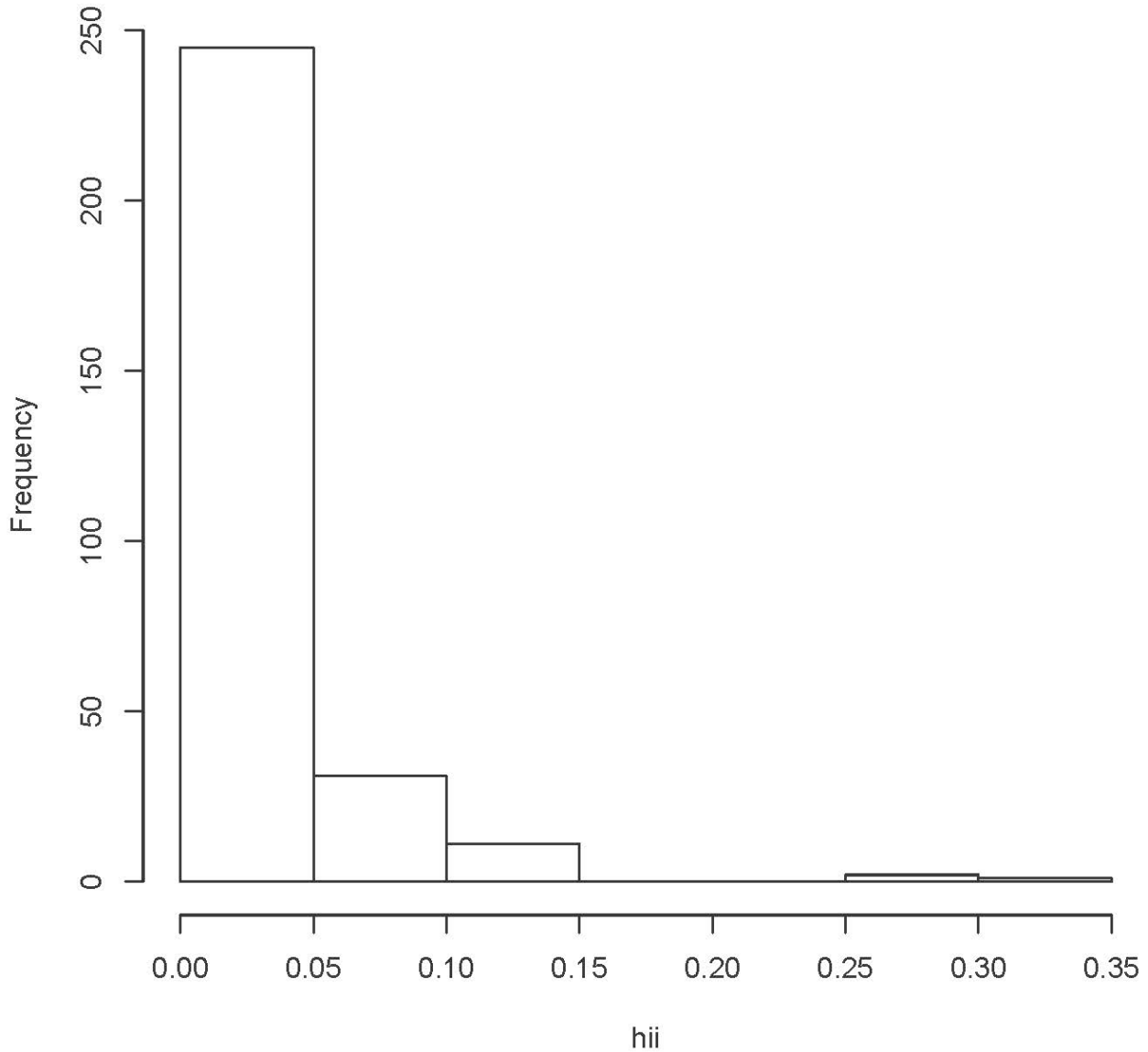
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-26.62766	10.25376	-2.597	0.009905	**
courseElite	-8.92387	5.40848	-1.650	0.100069	
courseMainstrm	-5.78575	4.61603	-1.253	0.211104	
precalcl	2.03123	0.59845	3.394	0.000788	***
calcl	0.77213	0.40292	1.916	0.056337	.
hsgpa	0.31636	0.09886	3.200	0.001532	**
hscalcl	0.58927	0.09270	6.357	8.33e-10	***
hsengl	0.10509	0.11304	0.930	0.353356	
frstlangOther	5.81009	2.24539	2.588	0.010170	*
sexM	-1.35617	1.82179	-0.744	0.457250	

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.43 on 280 degrees of freedom  
Multiple R-squared: 0.3996, Adjusted R-squared: 0.3803  
F-statistic: 20.71 on 9 and 280 DF, p-value: < 2.2e-16

```
>
> ##### Influence Measures #####
>
> # Leverage
> hii = hatvalues(fullmodell)
> summary(hii); hist(hii)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.008746 0.017470 0.025360 0.034480 0.039550 0.319500
```

**Histogram of hii**



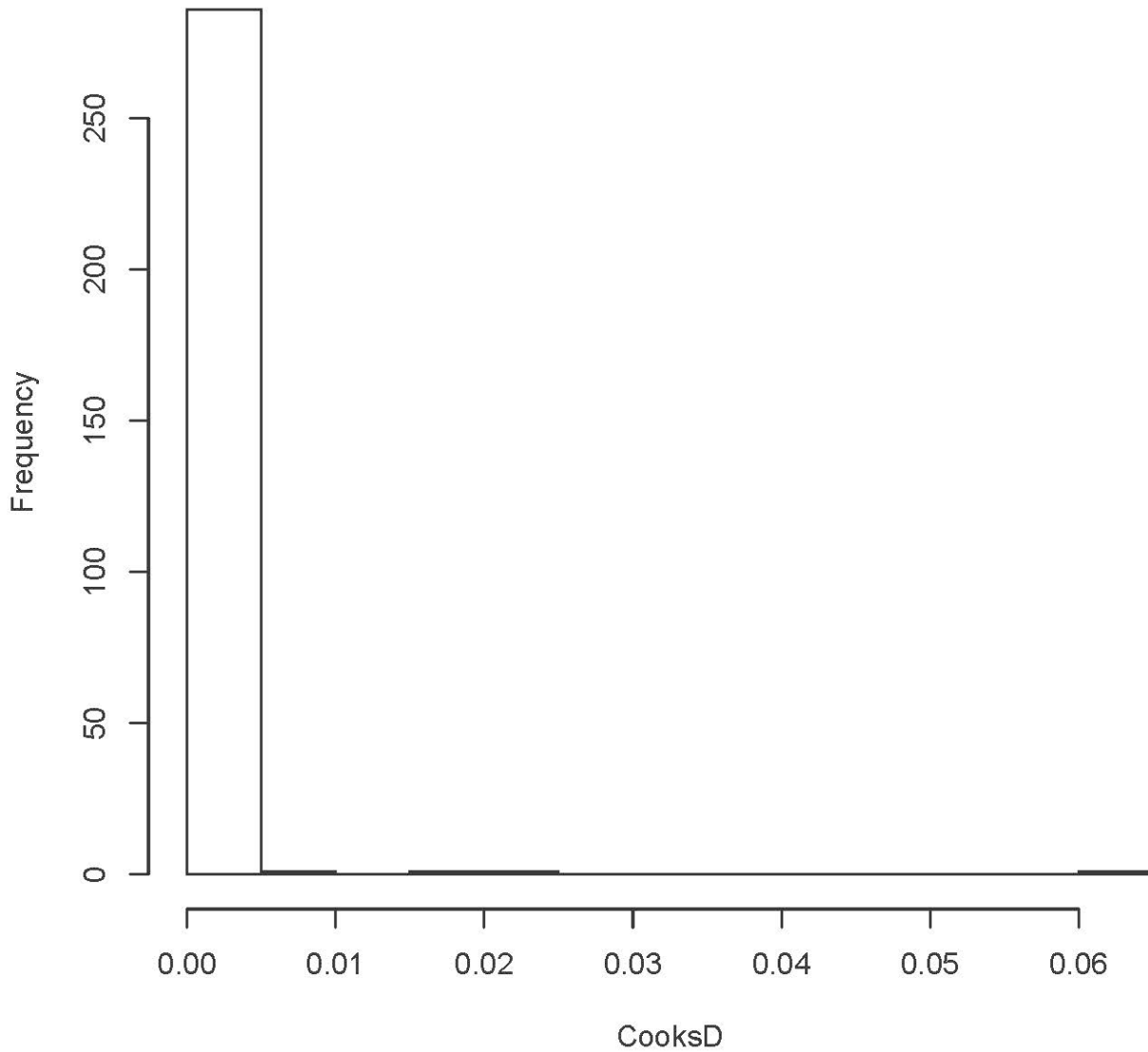
```
> hii[hii>0.2]
 116    127    133
0.2837035 0.2743562 0.3195286
>
```

```

> hii[hii>0.2]
      116      127      133
0.2837035 0.2743562 0.3195286
>
> # Cook's Distance
> n = length(frstlang)
> k = length(coefficients(fullmodell))
> ei = residuals(fullmodell)
> s = summary(fullmodell)$sigma
> CooksD = ei^2/((k+1)*s^2) * (hii/(1-hii))^2
> summary(CooksD); hist(CooksD)
      Min.      1st Qu.      Median      Mean      3rd Qu.      Max.
0.000e+00 3.810e-06 2.028e-05 4.847e-04 9.199e-05 6.067e-02

```

**Histogram of CooksD**





```

> 4/n # Rule of thumb says watch out if Cook's D > 4/n
[1] 0.0137931
> CooksD[CooksD>0.01] # Based on histogram
      116      127      133
0.06067450 0.02046928 0.01675851
> hii[hii>0.2] # Again, for comparison
      116      127      133
0.2837035 0.2743562 0.3195286

>
> math[c(116,127,133),] # These rows, all the columns
      course precalc calc hsgpa hscalcalc hsengl ucalc frstlang sex
235 Mainstrm      4      2      0      60      82      51 English F
256 Mainstrm      3      3      0      65      74      46 Other F
263 Mainstrm      7      7      0      91      62      65 Other M
> sort(hsgpa)
 [1]  0.0  0.0  0.0 66.0 66.0 69.0 69.8 70.0 70.5 70.7 70.7 71.8 72.2 72.3 72.3
 [16] 72.3 72.5 72.5 72.5 72.5 72.8 73.0 73.0 73.2 73.3 73.3 73.5 73.5 73.5 73.7
 [31] 73.7 73.8 73.8 74.0 74.0 74.2 74.3 74.3 74.5 74.5 74.5 74.5 74.5 74.7 74.7
 [46] 74.8 74.8 74.8 75.0 75.0 75.2 75.2 75.2 75.2 75.3 75.5 75.7 75.7 75.8 75.8
 [61] 75.8 76.0 76.0 76.2 76.2 76.2 76.2 76.3 76.3 76.3 76.5 76.7 76.7 76.7 76.8
 [76] 76.8 76.8 77.0 77.0 77.0 77.2 77.2 77.2 77.3 77.3 77.3 77.5 77.5 77.5 77.5
 [91] 77.5 77.5 77.7 77.7 77.8 78.0 78.0 78.0 78.0 78.0 78.0 78.2 78.2 78.2 78.2
[106] 78.2 78.3 78.3 78.3 78.3 78.3 78.3 78.3 78.3 78.3 78.5 78.5 78.5 78.5 78.7
[121] 78.7 78.7 78.8 78.8 79.0 79.0 79.3 79.3 79.3 79.3 79.5 79.5 79.5 79.5 79.7
[136] 79.7 79.8 79.8 80.0 80.0 80.0 80.0 80.0 80.0 80.0 80.0 80.2 80.2 80.3 80.3
[151] 80.5 80.5 80.5 80.7 80.7 80.7 80.7 80.8 80.8 80.8 81.0 81.0 81.0 81.0 81.0
[166] 81.0 81.0 81.0 81.2 81.2 81.2 81.2 81.3 81.3 81.5 81.7 81.7 81.7 81.7 81.8
[181] 81.8 81.8 81.8 82.0 82.0 82.0 82.0 82.0 82.0 82.2 82.2 82.2 82.2 82.3 82.3
[196] 82.5 82.5 82.7 82.7 82.8 82.8 83.0 83.0 83.0 83.2 83.3 83.3 83.3 83.5 83.7
[211] 83.8 83.8 84.0 84.0 84.0 84.2 84.3 84.3 84.5 84.7 84.7 84.7 84.7 84.8 85.2
[226] 85.3 85.8 85.8 86.2 86.3 86.5 86.5 86.7 86.7 86.7 86.7 86.8 86.8 86.8 86.8
[241] 87.0 87.3 87.3 87.3 87.8 88.0 88.2 88.3 88.3 88.3 88.5 88.5 88.5 88.7 88.7
[256] 88.7 89.0 89.2 90.2 90.3 90.3 90.5 90.5 90.8 91.0 91.0 91.0 91.5 91.5 91.5
[271] 91.5 91.8 92.3 92.3 92.3 92.3 92.5 92.5 92.7 93.2 93.3 93.5 93.5 93.7 93.7
[286] 94.0 94.3 95.0 95.8 96.2

>
> # Do the residuals reveal this?
> stdelres = rstudent(fullmodell)
> n = length(ucalc)
> a = 0.05/n
> critval = qt(1-a/2,fullmodell$df-1) # df = n-k-2
> critval # 3.807824
[1] 3.807824
> stdelres[abs(stdelres)>critval]
      86
-3.819437
> math[86,]
      course precalc calc hsgpa hscalcalc hsengl ucalc frstlang sex
176 Elite      3      6 69.8      76      60      1 Other M

```

```

> # Fix those three zeros (just locally) and re-fit the model
> hsgpa[hsgpa==0] = NA

> fullmodel2 = lm(ucalc ~ course + precalc + calc + hsgpa + hscalcal + hsengl +
frstlang + sex)
> summary(fullmodel2)

Call:
lm(formula = ucalc ~ course + precalc + calc + hsgpa + hscalcal +
hsengl + frstlang + sex)

Residuals:
    Min       1Q   Median       3Q      Max
-45.637  -6.557   1.183   8.517  30.945

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -66.7948    11.5843  -5.766 2.16e-08 ***
courseElite    -8.3597     5.0888  -1.643 0.10156
courseMainstrm -5.6157     4.3396  -1.294 0.19672
precalc        1.7374     0.5657   3.071 0.00235 **
calc           0.7117     0.3789   1.878 0.06140 .
hsgpa          1.5717     0.2211   7.110 9.89e-12 ***
hscalcal       0.2513     0.1033   2.433 0.01560 *
hsengl        -0.3273     0.1264  -2.590 0.01011 *
frstlangOther  4.4895     2.1334   2.104 0.03625 *
sexM          -1.6280     1.7155  -0.949 0.34344
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.56 on 277 degrees of freedom
(3 observations deleted due to missingness)
Multiple R-squared:  0.4737, Adjusted R-squared:  0.4566
F-statistic: 27.7 on 9 and 277 DF, p-value: < 2.2e-16

>
> # We are not done, but this illustrates the role of influence diagnostics
> # in data cleaning.

```

---

This document was prepared by [Jerry Brunner](#), University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License:

[http://creativecommons.org/licenses/by-sa/3.0/deed.en\\_US](http://creativecommons.org/licenses/by-sa/3.0/deed.en_US). Use any part of it as you like and share the result freely. The Open Office document is available from the course website at <http://www.utstat.toronto.edu/~brunner/oldclass/302f17>