Diary of a linear regression analysis

This handout describes a restricted analysis of a data set on autmotive fuel (i.e. gasoline) consumption across the United States in the early 1970's. The goal of the analysis is to find out what factors influence per-capita fuel consumption.

The format of this handout is a somewhat cleaned-up **data analysis diary**. The level of detail allows me to illustrate typical data analysis thinking, and to point out ways of analyzing the data inspired by our class discussion of the basic formulas and tools of regression analysis.

In a **data analysis report**, I would expect to see a shorter version of this handout as a *technical appendix*. The order of the material, the dominance of technical detail over substantive problem-solving, and the digressions into arcana of Splus and regression analysis, are *totally wrong* for the front matter, body or conclusions of a data analysis report.

1 The Fuel Consumption Data of Weisberg

This is a classic data set from S. Weisberg (1985). *Applied linear regression, 2nd Ed.* New York: Wiley, page 35. The data has nine columns,

```
(State) = Two-Letter State Code
  Pop   = Population in each state
  Tax   = 1971 Gas tax, cents per gallon
  Nlic  = 1971 thousands of drivers
  Inc   = 1972 per capita income
  Road  = 1971 miles of primary highways
  Fuelc = fuel consumption, millions of gallons
  Dlic  = 100*Nlic/Pop = proportion of licensed drivers
  Fuel  = 1000*Fuelc/Pop = per-capita fuel consumption
```

but I have arranged the data file `fuel.dat` so that the the "State" column will be read as the row-names by Splus's `read.table()` function, and the variable names are in the first row of the data file.

This data can be read as follows:

```
> fuel _ read.table("fuel.dat",header=T)
> fuel
     Pop   Tax  Nlic   Inc   Road Fuelc Dlic Fuel
ME  1029  9.00   540 3.571  1.976   557 52.5  541
NH   771  9.00   441 4.092  1.250   404 57.2  524
VT   462  9.00   268 3.865  1.586   259 58.0  561
MA  5787  7.50  3060 4.870  2.351  2396 52.9  414
RI   968  8.00   527 4.399  0.431   397 54.4  410
CN  3082 10.00  1760 5.342  1.333  1408 57.1  457
NY 18366  8.00  8278 5.319 11.868  6312 45.1  344
NJ  7367  8.00  4074 5.126  2.138  3439 55.3  467
PA 11926  8.00  6312 4.447  8.577  5528 52.9  464
OH 10783  7.00  5948 4.512  8.507  5375 55.2  498
IN  5291  8.00  2804 4.391  5.939  3068 53.0  580
IL 11251  7.50  5903 5.126 14.186  5301 52.5  471
MI  9082  7.00  5213 4.817  6.930  4768 57.4  525
WI  4520  7.00  2465 4.207  6.580  2294 54.5  508
MN  3896  7.00  2368 4.332  8.159  2204 60.8  566
IA  2883  7.00  1689 4.318 10.340  1830 58.6  635
```



```
MO  4753  7.00  2719 4.206  8.508  2865 57.2  603
ND   632  7.00   341 3.718  4.725   451 54.0  714
SD   579  7.00   419 4.716  5.915   501 72.4  865
NE  1525  8.50  1033 4.341  6.010   976 67.7  640
KS  2258  7.00  1496 4.593  7.834  1466 66.3  649
DE   565  8.00   340 4.983  0.602   305 60.2  540
MD  4056  9.00  2073 4.897  2.449  1883 51.1  464
VA  4764  9.00  2463 4.258  4.686  2604 51.7  547
WV  1781  8.50   982 4.574  2.619   819 55.1  460
NC  5214  9.00  2835 3.721  4.746  2953 54.4  566
SC  2665  8.00  1460 3.448  5.399  1537 54.8  577
GA  4720  7.50  2731 3.846  9.061  2979 57.9  631
FL  7259  8.00  4084 4.188  5.975  4169 56.3  574
KY  3299  9.00  1626 3.601  4.650  1761 49.3  534
TN  4031  7.00  2088 3.640  6.905  2301 51.8  571
AL  3510  7.00  1801 3.333  6.594  1946 51.3  554
MS  2263  8.00  1309 3.063  6.524  1306 57.8  577
AR  1978  7.50  1081 3.357  4.121  1242 54.7  628
LA  3720  8.00  1813 3.528  3.495  1812 48.7  487
OK  2634  6.58  1657 3.802  7.834  1695 62.9  644
TX 11649  5.00  6595 4.045 17.782  7451 56.6  640
MT   719  7.00   421 3.897  6.385   506 58.6  704
ID   756  8.50   501 3.635  3.274   490 66.3  648
WY   345  7.00   232 4.345  3.905   334 67.2  968
CO  2357  7.00  1475 4.449  4.639  1384 62.6  587
NM  1065  7.00   600 3.656  3.985   744 56.3  699
AZ  1945  7.00  1173 4.300  3.635  1230 60.3  632
UT  1126  7.00   572 3.745  2.611   666 50.8  591
NV   527  6.00   354 5.215  2.302   412 67.2  782
WN  3443  9.00  1966 4.476  3.942  1757 57.1  510
OR  2182  7.00  1360 4.296  4.083  1331 62.3  610
CA 20468  7.00 12130 5.002  9.794 10730 59.3  524
AK   302  8.00   136 5.162  3.246   166 45.2  551
HI   770  5.00   499 4.995  0.602   265 64.8  345
```

Our main task is to find out what variables influence per-capita fuel consumtion across the states.

For this analysis we will look at a restricted set of variables:

```
The variables we are interested in are:
Y  = Fuel = 1000*(fuel consumption per capita in each state)
X1 = Dlic = 100*(proportion of licenced drivers in the state)
X2 = Tax  = (Gas tax, cents per gallon, in the state)
```

We begin with some simple summaries

```
> fuel3 _ fuel[,c("Fuel","Dlic","Tax")]
>
> summary(fuel3)
      Fuel            Dlic            Tax
   Min.:344.0      Min.:45.10      Min.: 5.000
 1st Qu.:508.5   1st Qu.:52.92   1st Qu.: 7.000
  Median:566.0    Median:56.45    Median: 7.500
    Mean:571.6      Mean:56.95      Mean: 7.622
 3rd Qu.:631.8   3rd Qu.:59.98   3rd Qu.: 8.000
    Max.:968.0      Max.:72.40      Max.:10.000
```

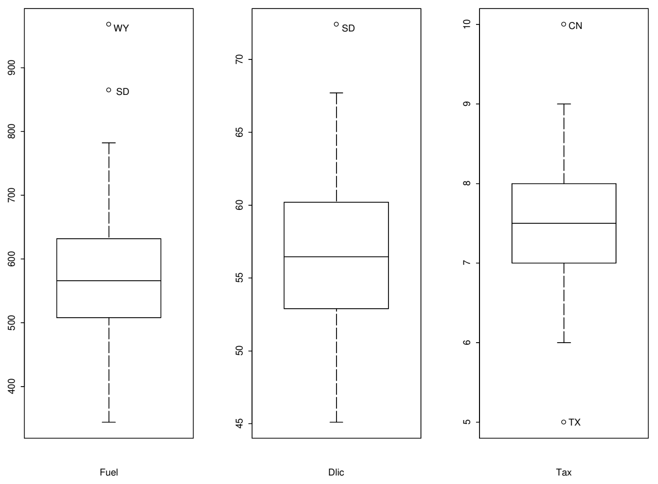

Figure 1: Boxplots with some initial guesses about outliers.

This can be elucidated with boxplots, as in Figure 1. The first boxplot was created with

```
> graphsheet() # On our UNIX workstations, use ‘‘motif()’’ here.
> par(mfrow=c(1,3))
> boxplot(fuel3$Fuel,style.bxp="old",xlab="Fuel")
> pts _ locator()
> pts
$x: [1] 49.58816 50.39790
$y: [1] 864.7793 967.6954
> identify(x=rep(50,length=50),y=fuel3$Fuel,pts=pts,labels=dimnames(fuel3)[[1]])
[1] 19 40
```

and similarly for the other two, substituting “Dlic” and “Tax” for “Fuel”. Note that some care is required in using the “locator” and “identify” functions to identify points. See `help()` on these functions for more details. According to the boxplot criteria, WY and SD are high outliers in per-capita `Fuel` consumption, SD is a high outlier in proportion of drivers' licences (`Dlic`), and CN and TX are respectively high and low outliers on `Tax` per gallon. (Some other outliers were covered up by those identified here; see next plot.)

We can also look at the two-way relationships between the variables using

```
> pairs(fuel3)
```

but I will be slightly fancier and show “yet another way” to plot the pairs so that we can identify interesting points

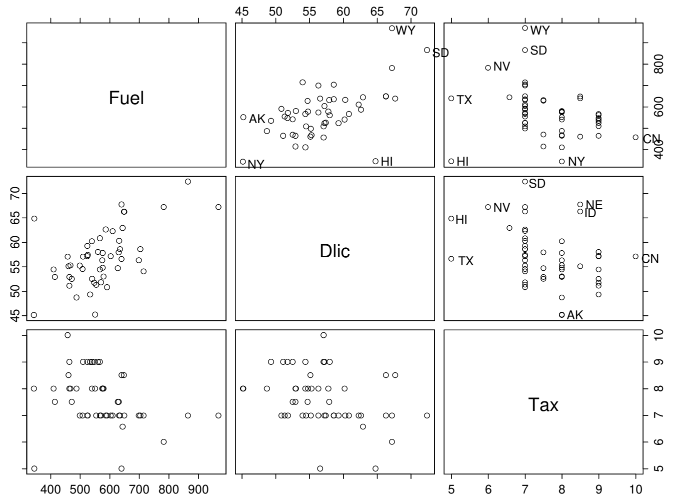

Figure 2: `pairs()` plot of the three fuel variable with outlying states labelled. See text for details.

```
> pairs(fuel3, panel=function(x,y) {
+  plot(x,y)
+  pts_locator()
+  identify(x,y,labels=dimnames(fuel3)[[1]],pts=pts)
+ }
+ )
Warning in axes: No room for xlab
Warning in axes: No room for ylab
Warning in axes: No room for X-axis
Warning in axes: No room for Y-axis
.
.
.
Warning in axes: No room for X-axis
Warning in axes: No room for Y-axis
```

Ignore all the warning messages (there are a lot). To see how I did this, look at the `help()` for `pairs()`, `identify()` and `locator()` functions. The result is shown in Figure 2. We can see from Figure 2 that

- There is a fairly clear increasing relationship between `Fuel` and `Dlic`;
- There is a less pronounced decreasing relationship between `Fuel` and `Tax`;
- There is a similar decreasing relationship between `Dlic` and `Tax`;

- `Tax` is clearly fairly discrete; we can see the different values of `Tax` by saying `sort(unique(fuel$Tax))` which tells us `[1] 5.00 6.00 6.58 7.00 7.50 8.00 8.50 9.00 10.00`.
- NV, WY and SD seem to be the high per-capita fuel consumers;
- NV, SD, ID, NE seem to have the highest proportion of drivers;
- HI is a low outlier (along with TX, which we identified with the boxplots above) on `Tax`. In fact, HI seems to be far from the center of the data in every scatter plot: low on `Fuel`, low on `Tax` and high on `Dlic`;
- AK and NY seem to have the smallest proportion of drivers, though the proportion was not low enough to be identified as outlying by the boxplots above;
- HI seems to be well away from any reasonable regression line (use your eye!) in the `Fuel by Dlic` and `Fuel by Tax` plots.

From these plots it seems reasonable to believe that `Y=Fuel` is related to both `X1=Dlic` and `X2=Tax`, but there is a little more to the story. It also appears that `X1` *and* `X2` *are related to each other*; one conclusion would be that less tax implies more drivers.

1.1 Some simple regressions

An obvious approach to this problem is to try several simple linear regressions, e.g.

```
> mod1 <- lm( Fuel ~ Dlic, data=fuel3)
> summary(mod1)

Call: lm(formula = Fuel ~ Dlic, data = fuel3)
Residuals:
    Min      1Q   Median     3Q   Max
 -316.6 -53.09 -0.07617 49.71 278.9

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  -81.3028  132.7597    -0.6124    0.5432
       Dlic   11.4644    2.3193     4.9430    0.0000

Residual standard error: 94.1 on 48 degrees of freedom
Multiple R-Squared: 0.3373
F-statistic: 24.43 on 1 and 48 degrees of freedom, the p-value is 9.787e-006

Correlation of Coefficients:
     (Intercept)
Dlic -0.995
> anova(mod1)
Analysis of Variance Table

Response: Fuel

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value         Pr(F)
     Dlic  1  216363.7 216363.7 24.43303 9.787074e-006
Residuals 48  425058.1   8855.4
```


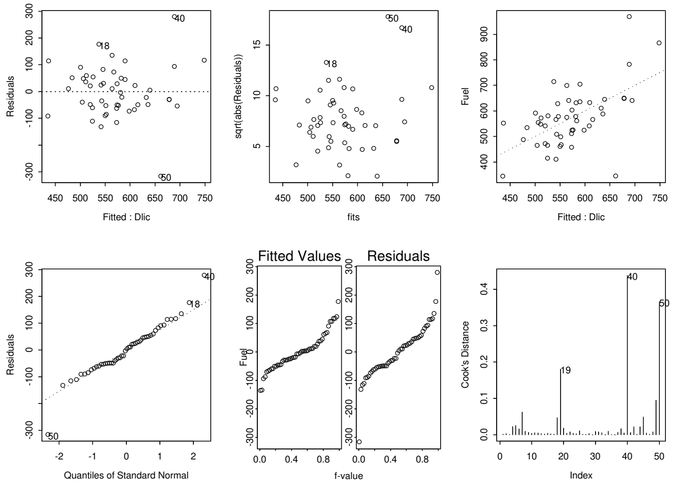

Figure 3: Standard Splus diagnostic plots for `lm(formula = Fuel ~ Dlic, data = fuel3)`.

```
> par(mfrow=c(2,3))
> plot(mod1)
```

Roughly 1/3 of the variability in fuel is explained by DLIC. Now consider

```
> mod2 <- lm( Fuel ~ Tax, data=fuel3)
> summary(mod2)

Call: lm(formula = Fuel ~ Tax, data = fuel3)
Residuals:
    Min     1Q Median    3Q   Max
 -314.3 -65.94   3.53 40.25 375.6

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  826.4201  120.5791     6.8538    0.0000
        Tax  -33.4313   15.6873    -2.1311    0.0382

Residual standard error: 110.5 on 48 degrees of freedom
Multiple R-Squared: 0.08644
F-statistic: 4.542 on 1 and 48 degrees of freedom, the p-value is 0.03823

Correlation of Coefficients:
    (Intercept)
Tax -0.9916
```

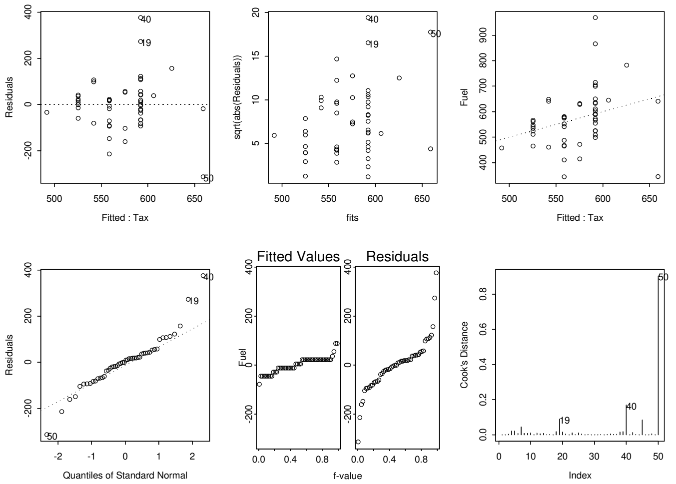

Figure 4: Standard Splus diagnostic plots for `lm(formula = Fuel ~ Tax, data = fuel3)`.

```
> anova(mod2)
Analysis of Variance Table

Response: Fuel

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
      Tax  1   55443.5 55443.48 4.541614 0.03822929
Residuals 48  585978.3 12207.88
> par(mfrow=c(2,3))
> plot(mod2)
```

1.2 Diagnostics for the simple regressions

It is useful to know

- which residuals are unusually large;
- which have large influence.

For the purpose of identifying the large residuals, the standard `plot`'s from Splus aren't all that useful. Now we know that

$$\text{Var}(e_i) = (1 - h_{ii})\sigma^2$$

where h_{ii} are the "diagonal entries in the hat matrix",

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

and we know

$$S^2 = \frac{RSS}{n-2} = \frac{1}{n-2}\sum_i e_i^2 = \frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2$$

is an unbiased estimate of σ^2. Thus, if the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, is correct,

$$e_i' = e_i / \sqrt{(1 - h_{ii})S^2}$$

should be approximately normally distributed [actually you could expect a t-distribution (why?), but there shouldn't be much difference for more than about 30 or 40 observations (why?)].

Splus has a function `lm.hat()` that computes the h_{ii}; and S^2 can be calculated from the `lm` objects we've created. I created an Splus function to do this,

```
stdres <- function(fitted.lm) {
  # for other magic, try "names(mod1)" at the SPLUS prompt
  resids <- fitted.lm$residuals
  df     <- fitted.lm$df.residual
  S2E    <- sum(resids^2)/df
  hats   <- lm.hat(fitted.lm)
  stdres <- resids/sqrt((1-hats)*S2E)
  return(stdres)
}
```

and saved it in the file `stdres.splus`; we now `source()` the file and use the function.

```
> source("stdres.splus")
> par(mfrow=c(2,2))

> stdres(mod1)
       ME         NH         VT        MA       RI        CN        NY         NJ
 0.220555 -0.5417093 -0.2430562 -1.199436 -1.42373 -1.248615 -1.030719 -0.9205223
         PA         OH        IN         IL         MI         WI         MN
 -0.6599567 -0.5752135 0.5791204 -0.5355216 -0.5556121 -0.3818871 -0.5363512
        IA        MO       ND       SD         NE         KS         DE
 0.4779457 0.3063355 1.896803 1.352195 -0.6109607 -0.3288067 -0.7415715
         MD        VA         WV        NC        SC       GA        FL        KY
 -0.4397645 0.3853643 -0.9713051 0.2542487 0.3230528 0.520896 0.1058008 0.5478986
        TN        AL          MS       AR        LA         OK        TX       MT
 0.6326037 0.5115133 -0.04661811 0.883747 0.1095159 0.04547381 0.7773767 1.219253
         ID       WY         CO       NM        AZ        UT      NV         WN
 -0.3398444 3.096259 -0.5352955 1.447795 0.2369529 0.9766585 1.03129 -0.6796806
         OR         CA       AK        HI
 -0.2483742 -0.8015073 1.280978 -3.465272
> cbind(Fitted=fitted(mod1),fuel3,stdres=stdres(mod1))[abs(stdres(mod1))>2,]
     Fitted Fuel Dlic Tax    stdres
WY 689.1076  968 67.2   7  3.096259
HI 661.5929  345 64.8   5 -3.465272

> plot(fitted(mod1),stdres(mod1))
```

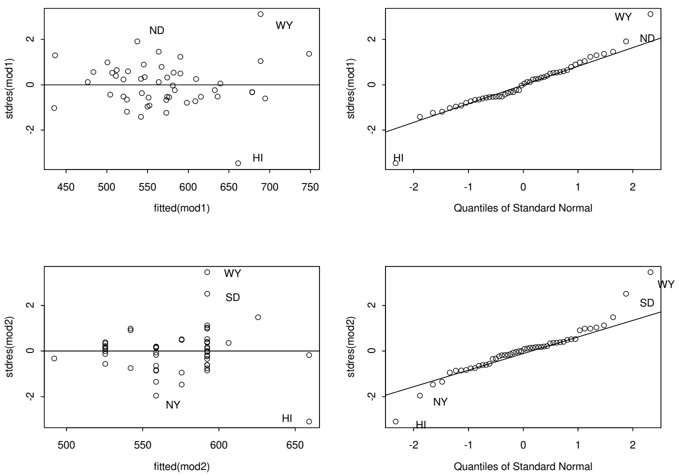

Figure 5: Standardized residuals for the two simple regressions, `mod1: Fuel ~ Dlic` and `mod2: Fuel ~ Tax`.

```
> abline(0,0)
> identify(fitted(mod1),stdres(mod1),labels=dimnames(fuel3)[[1]])
[1] 50 40 18
> qqnorm(stdres(mod1))
> qqline(stdres(mod1))
> identify(qnorm(rank(stdres(mod1))/50.5),stdres(mod1),labels=dimnames(fuel3)[[1]])
[1] 50 18 40
```

```
> stdres(mod2)
        ME          NH        VT        MA       RI         CN        NY
 0.1442054 -0.01434691 0.3307376 -1.478439 -1.36397 -0.3414461 -1.968269
        NJ         PA         OH        IN         IL         MI         WI
 -0.8420766 -0.8695447 -0.8665143 0.1925551 -0.9572345 -0.6186789 -0.7747234
        MN         IA         MO       ND       SD        NE        KS         DE
 -0.2423363 0.3910209 0.09729004 1.116169 2.502211 0.9008245 0.5195281 -0.1736862
        MD        VA         WV        NC       SC        GA        FL
 -0.5739434 0.2001651 -0.7580495 0.3773706 0.165087 0.5057942 0.1376189
        KY         TN         AL         MS        AR         LA        OK
 0.07891918 -0.1964408 -0.3524853 0.165087 0.4783624 -0.6589559 0.3472718
        TX       MT        ID       WY          CO        NM        AZ
 -0.1900642 1.024378 0.9745522 3.447657 -0.04957538 0.9784826 0.3634836
         UT       NV         WN        OR        CA          AK        HI
 -0.01285903 1.468025 -0.1449194 0.1615437 -0.627858 -0.07296983 -3.100689
```



```
> cbind(Fitted=fitted(mod2),fuel3,stdres=stdres(mod2))[abs(stdres(mod2))>2,]
     Fitted Fuel Dlic Tax    stdres
SD 592.4009  865 72.4   7  2.502211
WY 592.4009  968 67.2   7  3.447657
HI 659.2635  345 64.8   5 -3.100689

> plot(fitted(mod2),stdres(mod2))
> abline(0,0)
> identify(fitted(mod2),stdres(mod2),labels=dimnames(fuel3)[[1]])
[1] 50 19 40  7
> qqnorm(stdres(mod2))
> qqline(stdres(mod2))
> identify(qnorm(rank(stdres(mod2))/50.5),stdres(mod2),labels=dimnames(fuel3)[[1]])
[1] 50 40 19  7
```

A plot of these standardized residuals is in Figure 5. WY, HI show up as poorly fitted in both regressions, and SD does not do well in with `mod2: Fuel ~ Tax`.

Similarly, it is useful to know which observations have large "influence". What does this mean? Recall that the equation for $\hat{\beta}_1$ says now for example that

$$\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

so that, keeping SXX constant,

$$\frac{\partial \hat{\beta}_1}{\partial (y_i - \bar{y})} \propto (x_i - \bar{x})$$

i.e. the influence of a change in $y_i - \bar{y}$ on the slope of the regression line is proportional to the size of $x_i - \bar{x}$. Since h_{ii} is sensitive to $x_i - \bar{x}$, h_{ii} is a measure of this influence. (A better measure, called "Cook's distance" is also sensitive to how much $y_i - \bar{y}$ has "pulled" the regression line off-target.)

We can just use **`lm.hat()`** directly to list (and plot) the h_{ii} values. See also the function **`lm.influence()`**.

```
> par(mfrow=c(1,2))
> lm.hat(mod1)
         ME         NH         VT         MA         RI         CN        NY
 0.03204015 0.02003736 0.02066718 0.02997379 0.02395624 0.02001331 0.1053306
         NJ         PA         OH         IN         IL         MI         WI
 0.02165784 0.02997379 0.02186462 0.02948758 0.03204015 0.02012192 0.02365227
        MN         IA         MO         ND        SD         NE         KS
 0.0289948 0.02164982 0.02003736 0.02529364 0.1649659 0.09017408 0.07308341
         DE        MD         VA         WV         NC         SC         GA
 0.02640846 0.0408032 0.03675602 0.02208355 0.02395624 0.02281323 0.02054593
         FL         KY         TN         AL         MS         AR         LA
 0.02025824 0.05556897 0.03612401 0.03940554 0.02043683 0.02308076 0.06136565
         OK         TX         MT         ID         WY         CO         NM
 0.04149133 0.02007527 0.02164982 0.07308341 0.08379691 0.03937808 0.02025824
         AZ        UT         NV         WN         OR         CA        AK
 0.02680914 0.0429908 0.08379691 0.02001331 0.03737418 0.02334902 0.1038967
         HI
 0.05741445
> plot(lm.hat(mod1))
> identify(1:50,lm.hat(mod1),labels=dimnames(fuel3)[[1]])
[1]  7 19 20 49
```

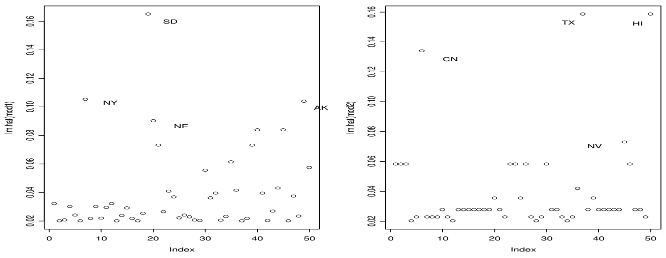

Figure 6: Influence measures (h_{ii}) for the two simple regressions, `mod1: Fuel ~ Dlic` and `mod2: Fuel ~ Tax`.

```
> lm.hat(mod2)
        ME         NH         VT         MA         RI        CN         NY
0.05830072 0.05830072 0.05830072 0.02029807 0.02288641 0.1340319 0.02288641
        NJ         PA         OH         IN         IL         MI         WI
0.02288641 0.02288641 0.02778894 0.02288641 0.02029807 0.02778894 0.02778894
        MN         IA         MO         ND         SD         NE         KS
0.02778894 0.02778894 0.02778894 0.02778894 0.02778894 0.03555396 0.02778894
        DE         MD         VA         WV         NC         SC         GA
0.02288641 0.05830072 0.05830072 0.03555396 0.05830072 0.02288641 0.02029807
        FL         KY         TN         AL         MS         AR         LA
0.02288641 0.05830072 0.02778894 0.02778894 0.02288641 0.02029807 0.02288641
        OK        TX         MT         ID         WY         CO         NM
0.04187048 0.1585445 0.02778894 0.03555396 0.02778894 0.02778894 0.02778894
        AZ         UT        NV         WN         OR         CA         AK
0.02778894 0.02778894 0.0730083 0.05830072 0.02778894 0.02778894 0.02288641
       HI
0.1585445
> plot(lm.hat(mod2))
> identify(1:50,lm.hat(mod2),labels=dimnames(fuel3)[[1]])
[1] 50 37  6 45
```

Figure 6 shows the h_{ii} plots for these two regressions. SD shows up in `mod1` as being influential, and HI shows up in `mod2`. Interestingly, CN, TX and perhaps NV appear to be influential in `mod2` also. NY and AK may be mildly influential in `mod1`. To summarize the two models,

Model	`mod1`	`mod2`
Formula	`Fuel ~ Dlic`	`Fuel ~ Tax`
R^2	0.3373	0.08644
Large e_i''s	WY, HI	SD, WY, HI
Large h_{ii}'s	SD, NY?, AK?	CN, TX, HI

Quite clearly the first model explains more of the variability in `Fuel`. To pursue either regression we would want to consider dropping the large-h_{ii} data points from the model (probably one-by-one, starting with the largest h_{ii} point first), and seeing how large the e_i's get for the remaining data

points[1].

We can improve things a bit more by combining "subject area knowledge" with the regression diagnostics: If we remove AK and HI so that we are restricted to the "continental 48 states" we get

```
> rbind(dimnames(fuel)[[1]],1:50)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
[1,] "ME" "NH" "VT" "MA" "RI" "CN" "NY" "NJ" "PA" "OH"  "IN"  "IL"  "MI"  "WI"
[2,] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10"  "11"  "12"  "13"  "14"
     [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26]
[1,] "MN"  "IA"  "MO"  "ND"  "SD"  "NE"  "KS"  "DE"  "MD"  "VA"  "WV"  "NC"
[2,] "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"  "23"  "24"  "25"  "26"
     [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37] [,38]
[1,] "SC"  "GA"  "FL"  "KY"  "TN"  "AL"  "MS"  "AR"  "LA"  "OK"  "TX"  "MT"
[2,] "27"  "28"  "29"  "30"  "31"  "32"  "33"  "34"  "35"  "36"  "37"  "38"
     [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48] [,49] [,50]
[1,] "ID"  "WY"  "CO"  "NM"  "AZ"  "UT"  "NV"  "WN"  "OR"  "CA"  "AK"  "HI"
[2,] "39"  "40"  "41"  "42"  "43"  "44"  "45"  "46"  "47"  "48"  "49"  "50"
> fuel48 _ fuel[-c(49,50),]
> mod1 _ lm(Fuel ~ Dlic,data=fuel48)
> summary(mod1)
[...]
Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept) -227.3091  121.8617    -1.8653    0.0685
       Dlic   14.0984    2.1268     6.6288    0.0000

Residual standard error: 80.88 on 46 degrees of freedom
Multiple R-Squared: 0.4886
F-statistic: 43.94 on 1 and 46 degrees of freedom, the p-value is 3.29e-008
[...]
> mod2 _ lm(Fuel ~ Tax,data=fuel48)
> summary(mod2)
[...]
Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  984.0076  119.6236     8.2259    0.0000
        Tax  -53.1063   15.4836    -3.4298    0.0013

Residual standard error: 100.9 on 46 degrees of freedom
Multiple R-Squared: 0.2037
F-statistic: 11.76 on 1 and 46 degrees of freedom, the p-value is 0.001285
[...]
```

Now the first regression explains almost 50% of the variability in `Fuel`, and the second explains about 20%. But, is that all the data has to say?

1.3 Multiple regression

Let's look at the residuals from "Fuel vs. Dlic" plotted against Tax.

[1] A complication is that removal of one large-influence point can reveal another moderately-large influence point whose influence was "hiding behind" the point you just removed. For example if CN, TX and HI are removed from `mod2`, the h_{ii} value of NV increases from 0.0730 to 0.1155 (try it!).

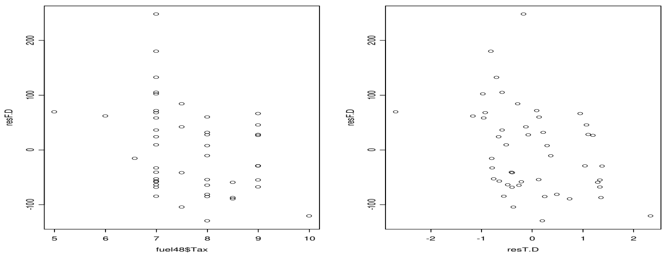

Figure 7: Predicting the residuals of Fuel vs. Dlic with Tax

```
> par(mfrow=c(1,2))
> resF.D _ residuals(lm(Fuel ~ Dlic, data=fuel48))
> plot(fuel48$Tax,resF.D)
> resT.D _ residuals(lm(Tax  ~ Dlic, data=fuel48))
> plot(resT.D,resF.D)
```

There is still some evidence that the residuals are related to `Tax`. A simple regression of the residuals confirms that there is indeed a relationship.

```
> resreg _ lm(resF.D ~ resT.D)
> summary(resreg)

Call: lm(formula = resF.D ~ resT.D)
Residuals:
    Min     1Q Median    3Q   Max
 -123.2 -60.17 -2.908 45.03 242.6

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.0000  10.8688     0.0000   1.0000
     resT.D -32.0753  12.0639    -2.6588   0.0108

Residual standard error: 75.3 on 46 degrees of freedom
Multiple R-Squared: 0.1332
F-statistic: 7.069 on 1 and 46 degrees of freedom, the p-value is 0.01075

Correlation of Coefficients:
       (Intercept)
resT.D 0
```

But these two simple regressions have not captured all of the (linear) variability in the data. Perhaps if we fit the model

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \text{error}$$

we can push R^2 up to 70% (50% + 20%)? Does that seem reasonable?

First let us note the correlations


```
> cor(fuel48[,c("Fuel", "Dlic", "Tax")])
           Fuel       Dlic        Tax
Fuel  1.0000000  0.6989654 -0.4512803
Dlic  0.6989654  1.0000000 -0.2880372
 Tax -0.4512803 -0.2880372  1.0000001
```

Now let's see the regression:

```
>  mod3 <- lm( Fuel ~ Dlic + Tax, data=fuel48) # in that order
> summary(mod3)

Call: lm(formula = Fuel ~ Dlic + Tax, data = fuel48)
Residuals:
    Min     1Q Median    3Q   Max
 -123.2 -60.17 -2.908 45.03 242.6

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  108.9709  171.7859     0.6343    0.5291
       Dlic   12.5149    2.0906     5.9862    0.0000
        Tax  -32.0753   12.1972    -2.6297    0.0117

Residual standard error: 76.13 on 45 degrees of freedom
Multiple R-Squared: 0.5567
F-statistic: 28.25 on 2 and 45 degrees of freedom, the p-value is 1.125e-008

Correlation of Coefficients:
     (Intercept)    Dlic
Dlic -0.8509
 Tax -0.7444      0.2880
> anova(mod3)
Analysis of Variance Table

Response: Fuel

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
Dlic       1  287448.0 287448.0 49.59155 0.00000001
Tax        1   40084.5  40084.5  6.91552 0.01165401
Residuals 45  260834.0   5796.3

> cbind(Fitted=fitted(mod3),fuel48[,c("Fuel", "Dlic", "Tax")],
+ stdres=stdres(mod3))[abs(stdres(mod3))>2,]
     Fitted Fuel Dlic Tax   stdres
ND 560.2461  714 54.0   7 2.065885
WY 725.4423  968 67.2   7 3.345335

> stem(lm.hat(mod3))

N = 48   Median = 0.0443331
Quartiles = 0.0324329, 0.0712345

Decimal point is 2 places to the left of the colon
```

```
    2 : 2345677
    3 : 112223344567789
    4 : 1446
    5 :
    6 : 12377779
    7 : 0023
    8 : 15
    9 : 3
   10 :
   11 :
   12 : 137
   13 :
   14 : 7

High: 0.1607706 0.1843257 0.2066504

> cbind(Fitted=fitted(mod3),fuel48[,c("Fuel", "Dlic", "Tax")],
+ stdres=stdres(mod3),hii=lm.hat(mod3))[lm.hat(mod3) > 0.16,]
     Fitted Fuel Dlic Tax     stdres       hii
CN 502.8162  457 57.1  10 -0.6569055 0.1607706
SD 790.5195  865 72.4   7  1.0831988 0.1843257
TX 656.9354  640 56.6   5 -0.2497392 0.2066504

> par(mfrow=c(2,2))
> plot(fitted(mod3),stdres(mod3))
> abline(0,0)
> identify(fitted(mod3),stdres(mod3),labels=dimnames(fuel48)[[1]])
[1] 40 18

> qqnorm(stdres(mod3))
> qqline(stdres(mod3))
> identify(qnorm(rank(stdres(mod3))/48.5),stdres(mod3),labels=dimnames(fuel48)[[1]])
[1] 18 40

> boxplot(lm.hat(mod3),style.bxp="old")
> pts _ locator()
> pts
$x: [1] 51.95419 52.50596 52.50596 52.50596
$y: [1] 0.1463142 0.1617925 0.1843221 0.2068516
> identify(rep(50,48),lm.hat(mod3),labels=dimnames(fuel48)[[1]],pts=pts)
[1] 20  6 19 37

> plot(lm.hat(mod3))
> identify(1:48,lm.hat(mod3),labels=dimnames(fuel48)[[1]],pts=locator())
[1]  7  6 20 19 37

> par(mfrow=c(2,3))
> plot(mod3)
> dimnames(fuel48)[[1]][c(5,18,19,40)]
[1] "RI" "ND" "SD" "WY"
```

What has happened? Why haven't things improved as much as I wanted? Why is R^2 not any bigger?

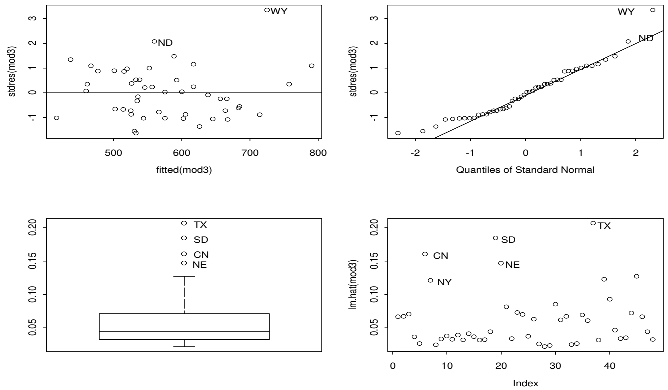

Figure 8: Standardized residuals and hii's for `lm(formula = Fuel ~ Dlic + Tax, data = fuel48)`.

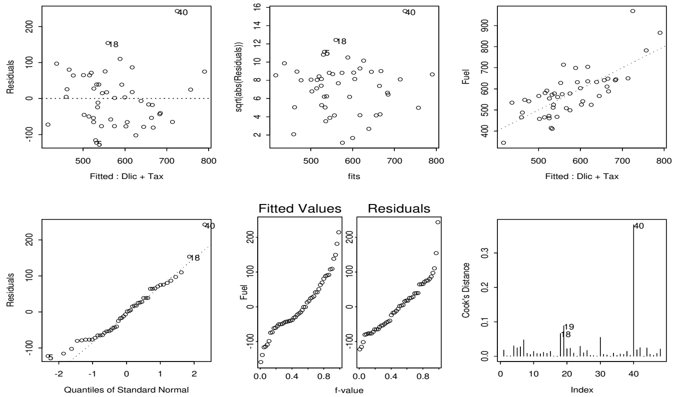

Figure 9: Splus's standard diagnistic plots for `lm(formula = Fuel ~ Dlic + Tax, data = fuel48)`.