Homework 5, Math 455: Due Monday, 03/26/2018

Alexander Van Roijen

March 26, 2018

Instructions: The homework assignment editing this LaTeX document. Download the LaTeX source from the class web page and study it to learn more about LaTeX. Replace the text with appropriate information. Run "pdflatex" on this document.

You will submit this assignment in two parts:

1. Print out the PDF file and bring it to class, and
2. Send an e-mail to:

   gang@math.binghamton.edu

   before class on the due date with two attachments:

   - The LaTeX source file, and
   - The generated PDF document.

Please complete the following:

1. Finish 6.1, 6.2,6.3,6.4, 6.5, 6.6 from the textbook. Submit your answers for ALL questions. NOTE: ALL graphs and images that are not deemed meaningful will not be included

1. We first check the constant variance assumption. At a glance, it seems to be fine, however you may be able to argue it has a non linear pattern.

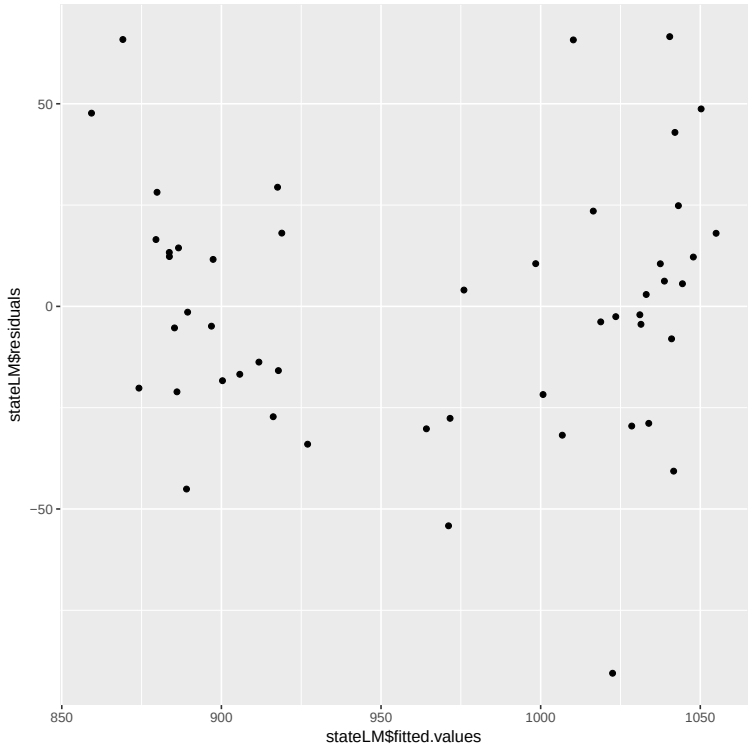

Figure 1: constant variance check on sat data

we then check our normality assumption using a qqplot. looks good still.
We check for large leverage points and find the following

```
> hats[which.max(abs(hats))]
Utah
0.2921128
```

further, looking at the following halfnormal graph for cooks distance, we again see Utah

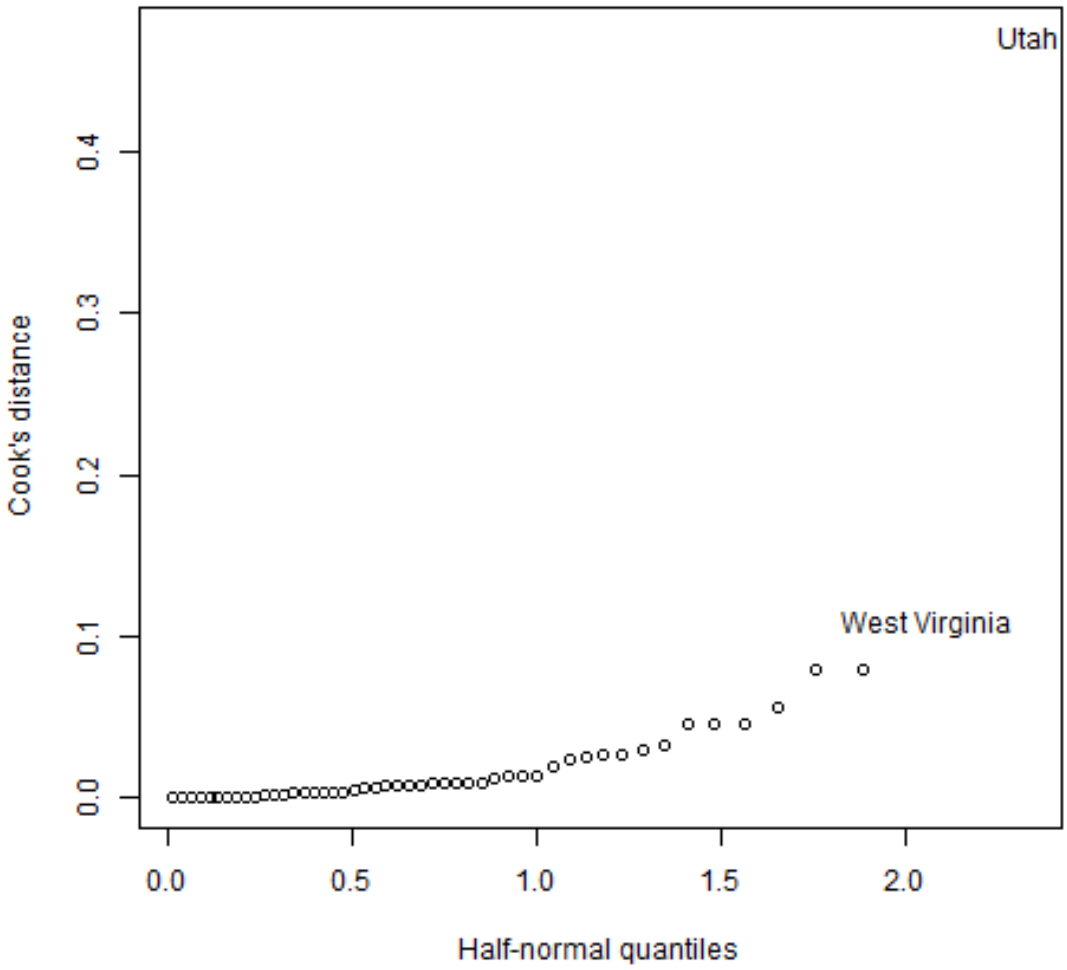

Figure 2: cooks distances from sat data

Taking a closer look:

```
> sat['Utah',]
expend ratio salary takers verbal math total
Utah  3.656  24.3 29.082      4    513  563  1076
```

We have a very high total, but a very low expenditure and a very low percentage of takers. This is giving a really high total for a state with very few takers. This outlier throws off

our model quite a bit. Removing it, we see the following difference in model values and performance

```
> stateLM2 = lm(total~expend+salary+ratio+takers,subset = satCook < max(satCook),sat)
> summary(stateLM)

Call:
lm(formula = y ~ expend + salary + ratio + takers, data = sat)

Residuals:
Min      1Q  Median      3Q     Max
-90.531 -20.855  -1.746  15.979  66.571

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
expend         4.4626    10.5465   0.423    0.674
salary         1.6379     2.3872   0.686    0.496
ratio         -3.6242     3.2154  -1.127    0.266
takers        -2.9045     0.2313 -12.559 2.61e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 32.7 on 45 degrees of freedom
Multiple R-squared:  0.8246,Adjusted R-squared:  0.809
F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16

> summary(stateLM2)

Call:
lm(formula = total ~ expend + salary + ratio + takers, data = sat,
```

```
subset = satCook < max(satCook))


Residuals:
Min      1Q  Median      3Q     Max
-92.118 -18.402   1.808  14.890  67.669


Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1093.8460    53.4226  20.475   <2e-16 ***
expend        -0.9427    10.1922  -0.092    0.927
salary         3.0964     2.3283   1.330    0.190
ratio         -7.6391     3.4279  -2.229    0.031 *
takers        -2.9308     0.2188 -13.397   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


Residual standard error: 30.9 on 44 degrees of freedom
Multiple R-squared:  0.8396,Adjusted R-squared:  0.825
F-statistic: 57.58 on 4 and 44 DF,  p-value: < 2.2e-16
```

We see that ratio becomes insignificant if we remove utah from this data set. We also see an improved R squared value.

I personally would reconsider removing utah, as I find some more information from utah might help quite a bit.

Note: summaries of data sets will be deprecated to key information from this point forward. In detail summaries can be found in the R code.

2. looking at our teen gambling data set, we immediately see that there is a non constant variance

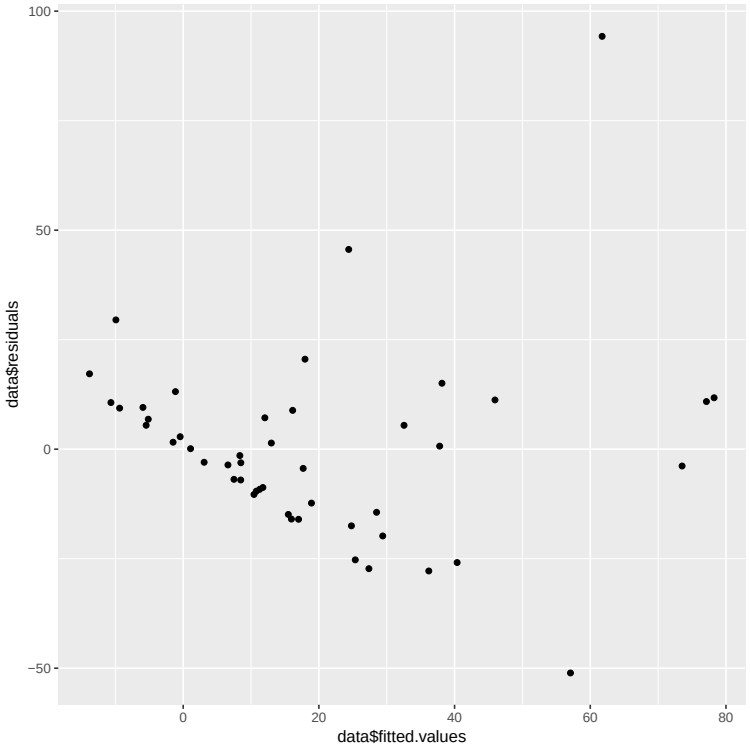

Figure 3: constant variance check on teen gamb data

we try to improve it by some transformations, but it still isnt quite satisfactory. The best is upon taking the square root of the response variable.

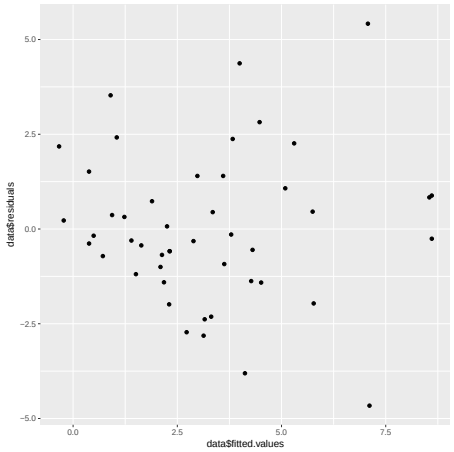

Figure 4: constant variance check on teen gamb data

However, we continue to identify any outliers and come across one instance, row 24, we

have a very high cooks distance

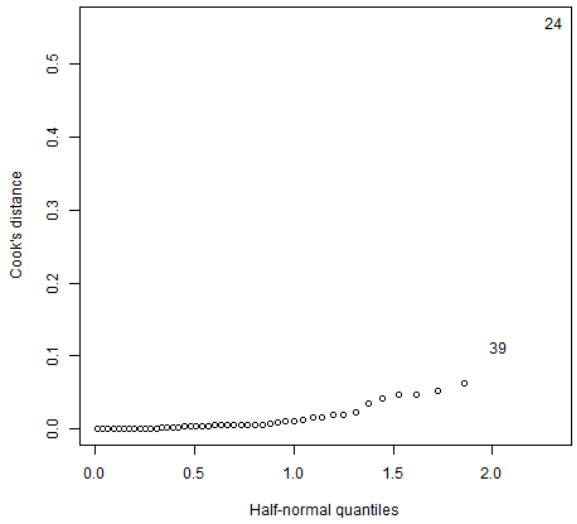

Figure 5: cooks distances from teen gamb data

we look at row 24, see it has a very high gamble value, and we then retry our model with this particular data point removed.

```
> teengamb[24,]
sex status income verbal gamble
24   0     27     10      4    156

> tmodel2 = lm(gamble~.,subset = weirdo< max(weirdo),teengamb)
> summary(tmodel)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.55565   17.19680   1.312   0.1968
sex         -22.11833    8.21111  -2.694   0.0101 *
status        0.05223    0.28111   0.186   0.8535
income        4.96198    1.02539   4.839 1.79e-05 ***
```

```
verbal       -2.95949    2.17215  -1.362   0.1803

Residual standard error: 22.69 on 42 degrees of freedom
Multiple R-squared:  0.5267,Adjusted R-squared:  0.4816
F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06

> summary(tmodel2)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.6306    12.9251   0.590   0.5582
sex         -16.2986     6.1335  -2.657   0.0112 *
status        0.1739     0.2083   0.835   0.4088
income        4.3312     0.7636   5.672 1.26e-06 ***
verbal       -1.8019     1.6137  -1.117   0.2707

Residual standard error: 16.74 on 41 degrees of freedom
Multiple R-squared:  0.5682,Adjusted R-squared:  0.526
F-statistic: 13.49 on 4 and 41 DF,  p-value: 4.225e-07
```

again, we see adjusted p values, coefficient values, and R squared.

3. Looking at the prostate data set, everything looks good in regards to structure, and normality assumptions. When we look at cooks distances however, we find quite a few values having abnormally high values.

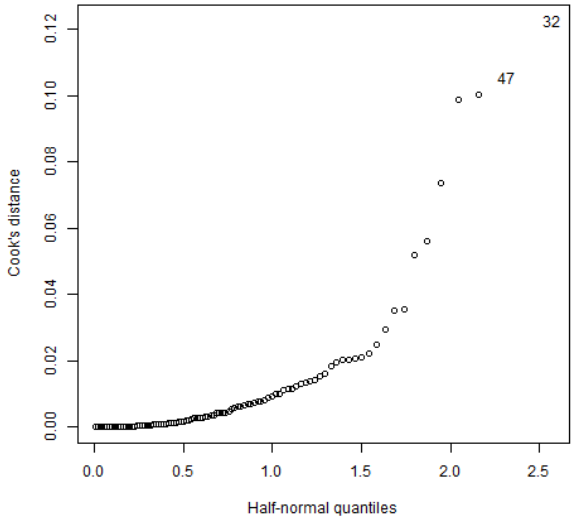

Figure 6: cooks distances from prostate data

we remove the highest point, 32, and see the below changes in our models.

```
> summary(pmodel)


Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886

Residual standard error: 0.7084 on 88 degrees of freedom
Multiple R-squared:  0.6548,Adjusted R-squared:  0.6234
```

```
F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16

> pmodel2=lm(lpsa~.,subset=cooksP<max(cooksP),prostate)
> summary(pmodel2)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.171863   1.328822   0.129  0.89739
lcavol       0.565333   0.088472   6.390 7.93e-09 ***
lweight      0.621663   0.202017   3.077  0.00279 **
age         -0.021271   0.011146  -1.908  0.05963 .
lbph         0.095590   0.058529   1.633  0.10604
svi          0.760423   0.242596   3.135  0.00235 **
lcp         -0.105987   0.090365  -1.173  0.24404
gleason      0.050688   0.156384   0.324  0.74662
pgg45        0.004468   0.004390   1.018  0.31155

Residual standard error: 0.7034 on 87 degrees of freedom
Multiple R-squared:  0.6629,Adjusted R-squared:  0.6319
F-statistic: 21.39 on 8 and 87 DF,  p-value: < 2.2e-16
```

Again, we similar changes. However, there were still more points that had high cooks distance values. We may consider looking at those points some more as well if we were doing extensive research with the data set.

4. Looking at the swiss data set, we are examining fertility measures of different provinces in switzerland. There is nothing too striking about this data set. Again we look at cooks distances and find that Porrentruy stands out.

```
> swiss['Porrentruy',]
Fertility Agriculture Examination Education Catholic Infant.Mortality
```

```
Porrentruy      76.1        35.3           9         7    90.57             26.6
```

this province has the highest infant mortality rate, which is likely why it is such an outlier.// Removing this from the data and adjusting a new model, we get teh following summaries

```
> smodel2 = lm(Fertility~.,subset=cooksS<max(cooksS),swiss)
> summary(smodel)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10

> summary(smodel2)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)      65.45554   10.16998   6.436 1.15e-07 ***
Agriculture      -0.21034    0.06859  -3.067  0.00387 **
Examination      -0.32278    0.24227  -1.332  0.19031
Education        -0.89506    0.17384  -5.149 7.36e-06 ***
Catholic          0.11269    0.03363   3.351  0.00177 **
Infant.Mortality  1.31567    0.37571   3.502  0.00115 **
```

```
Residual standard error: 6.794 on 40 degrees of freedom
Multiple R-squared:  0.7415,Adjusted R-squared:  0.7091
F-statistic: 22.94 on 5 and 40 DF,  p-value: 8.583e-11
```

We see another improved R squared and improved p values, particularly for agriculture. it is interesting to note the disparity of regions that are catholic. The following graph shows the divide.

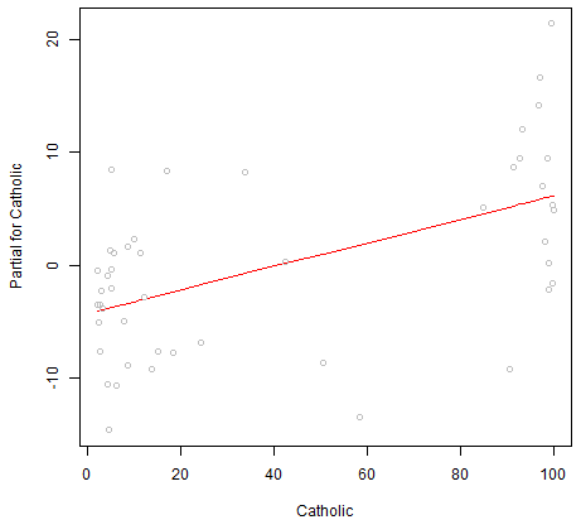

Figure 7: We see a seperation between catholic regions belows 30% and regions above 80%

It may be interesting to see if there is any difference between the two types of regions

5. Looking at the cheddar dataset, describing the taste of certain cheeses along with several parameters. We look for some outliers.

   We find there is nothing abnormal about the error, and there is only a few high leverages, but the entry that really sticks out is row number 15 due to its high studentized residual and cooks distance.

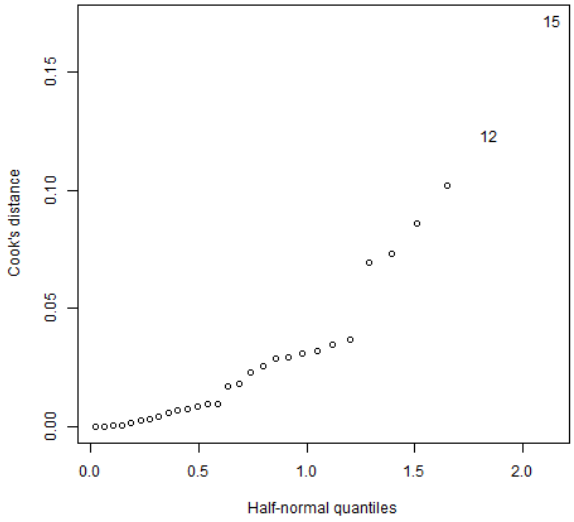

Figure 8: cooks distances from cheddar data

a closer look at fifteen reveals

```
> cheddar[15,]
taste Acetic   H2S Lactic
15   54.9  6.151 6.752   1.52
```

holding a very high taste value, but holding a low lactic value compared to other highly rated cheeses.

removing it from the model, we see the following performance change.

```
> cmodel2=lm(taste~.,subset=cooksC<max(cooksC),cheddar)
> summary(cmodel)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
Acetic        0.3277     4.4598   0.073  0.94198
```

```
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 *

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,Adjusted R-squared:  0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06

> summary(cmodel2)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.757     17.625  -1.007  0.32335
Acetic        -2.470      4.004  -0.617  0.54279
H2S            4.039      1.091   3.702  0.00106 **
Lactic        21.458      7.559   2.839  0.00886 **

Residual standard error: 8.847 on 25 degrees of freedom
Multiple R-squared:  0.7083,Adjusted R-squared:  0.6733
F-statistic: 20.24 on 3 and 25 DF,  p-value: 7.18e-07
```

Yet another large jump in r squared values and improved p values, particularly in H2S and lactic, which makes sense due to our previous outliers lactic value.

6. The happy data set, looks at how happy students are from the university of chicagos MBA program.
   the normality error assumptions check out, as in we have a seemingly constant variance, normally distributed, without any non linearity.
   looking at the studentized residuals, we notice entry 36 is quite high

```
> examineStudRes(hmodel)
36          19          16           7          38          37          20          23
```

```
-3.27601274 -1.67801954 -1.45766638 -1.27738113 -0.99473268 -0.96624585 -0.91144453 -0
```

with the accompanying diagram

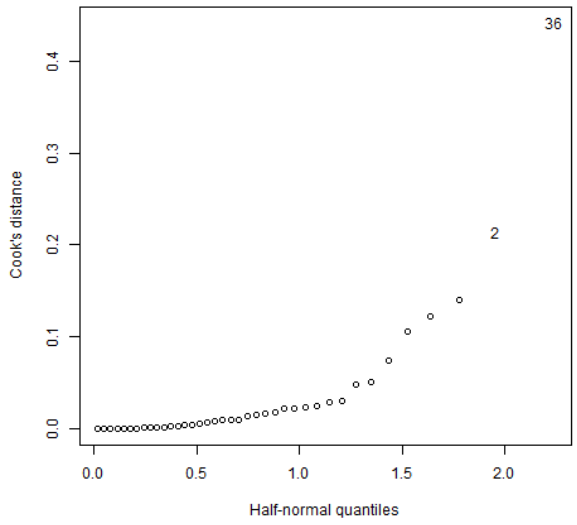

Figure 9: cooks distances from happy data set

we see that entry 36 is looking like quite the outlier, looking deeper.

```
> happy[36,]
happy money sex love work
36     2     0   0    2    2
```

very clear outlier here. There is some incorrect data input as it claims a household family income of 0 dollars. This could be an indication that they have no family, but for now it is too extreme. Removing the point and comparing models we get...

```
Call:
lm(formula = happy ~ ., data = happy)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -0.072081   0.852543  -0.085   0.9331
money        0.009578   0.005213   1.837   0.0749 .
sex         -0.149008   0.418525  -0.356   0.7240
love         1.919279   0.295451   6.496 1.97e-07 ***
work         0.476079   0.199389   2.388   0.0227 *

Residual standard error: 1.058 on 34 degrees of freedom
Multiple R-squared:  0.7102,Adjusted R-squared:  0.6761
F-statistic: 20.83 on 4 and 34 DF,  p-value: 9.364e-09

> summary(hmodel2)

Call:
lm(formula = happy ~ ., data = happy, subset = cooksH < max(cooksH))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.920460   0.810476   1.136    0.264
money        0.006680   0.004681   1.427    0.163
sex         -0.487967   0.383259  -1.273    0.212
love         1.953712   0.260722   7.493 1.29e-08 ***
work         0.305096   0.183392   1.664    0.106

Residual standard error: 0.9332 on 33 degrees of freedom
Multiple R-squared:  0.7347,Adjusted R-squared:  0.7026
F-statistic: 22.85 on 4 and 33 DF,  p-value: 4.064e-09
```

we see a minor increase in R squared values, and major improvements in p values except for money, which interestingly, went up after removing the term. This also made work no longer a powerful predictor.