liquid<-read.csv(file.choose(), header=TRUE)

#file name is Liquidity

attach(liquid)

summary(liquid)

      AVGT            VOLUME           NTRAN           PRICE

 Min.   : 0.590   Min.   : 0.658   Min.   :  999   Min.   :  9.125

 1st Qu.: 2.916   1st Qu.: 5.841   1st Qu.: 3090   1st Qu.: 24.562

 Median : 4.284   Median :11.556   Median : 5071   Median : 34.375

 Mean   : 5.441   Mean   :13.423   Mean   : 6436   Mean   : 38.796

 3rd Qu.: 7.010   3rd Qu.:17.509   3rd Qu.: 7414   3rd Qu.: 46.312

 Max.   :20.772   Max.   :64.572   Max.   :36420   Max.   :122.375

     SHARE             VALUE             DEBEQ              TIC

 Min.   :  6.736   Min.   : 0.1150   Min.   : 0.1850    AA    :  1

 1st Qu.: 28.402   1st Qu.: 0.9975   1st Qu.: 0.8005    AAL   :  1

 Median : 53.827   Median : 2.0650   Median : 1.1050    AEP   :  1

 Mean   : 94.731   Mean   : 4.1160   Mean   : 2.6973    AGE   :  1

 3rd Qu.:112.028   3rd Qu.: 3.9755   3rd Qu.: 1.7440    AHS   :  1

 Max.   :783.051   Max.   :75.4370   Max.   :53.6280    ALN   :  1

                                                       (Other):117

                   COMPANY

  ALEXANDER & ALEX SVCS:  1

  ALLEN GROUP INC      :  1

  ALUMINUM CO AMER     :  1

  AMERICAN ELEC PWR INC:  1

  AMERICAN HOSP SUPPLY :  1

  AMOCO CORP           :  1

 (Other)               :117

**model1=lm(VOLUME ~ PRICE + SHARE + VALUE)**

**summary(model1)**

Call:

lm(formula = VOLUME ~ PRICE + SHARE + VALUE)

Residuals:

    Min      1Q  Median      3Q     Max

-20.708  -4.179  -1.091   3.108  28.230

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  7.90916    1.60140   4.939 2.59e-06 \*\*\*

PRICE       -0.02224    0.03503  -0.635   0.5267

SHARE        0.05372    0.01034   5.194 8.63e-07 \*\*\*

VALUE        0.31271    0.16150   1.936   0.0552 .

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.722 on 119 degrees of freedom

Multiple R-squared:  0.6101,           Adjusted R-squared:  0.6003

F-statistic: 62.07 on 3 and 119 DF,  p-value: < 2.2e-16

step(model1)

Start:  AIC=472.66

VOLUME ~ PRICE + SHARE + VALUE

        Df Sum of Sq    RSS    AIC

- PRICE  1     18.22 5395.5 471.08

<none>               5377.3 472.66

- VALUE  1    169.41 5546.7 474.48

- SHARE  1   1218.82 6596.1 495.79

Step:  AIC=471.08

VOLUME ~ SHARE + VALUE

        Df Sum of Sq    RSS    AIC

<none>               5395.5 471.08

- VALUE  1    164.96 5560.5 472.78

- SHARE  1   1606.73 7002.2 501.14

Call:

lm(formula = VOLUME ~ SHARE + VALUE)

Coefficients:

(Intercept)        SHARE        VALUE

    7.03198      0.05638      0.25494

> # stepwise regression chooses a model with only SHARE and VALUE as X's

> # variable that's doing the worst job is PRICE, so we are going to build a new model and remove that.

> With backwards regression, you find the variable with the worst p-value and remove it and rerun the regression. That is PRICE here.

Error: unexpected symbol in "With backwards"

> model2 = lm(VOLUME ~ SHARE + VALUE)

> summary(model2)

Call:

lm(formula = VOLUME ~ SHARE + VALUE)

Residuals:

    Min      1Q  Median      3Q     Max

-20.782  -3.986  -1.695   3.328  28.087

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.031982   0.807719   8.706 2.02e-14 \*\*\*

SHARE       0.056385   0.009432   5.978 2.39e-08 \*\*\*

VALUE       0.254939   0.133097   1.915   0.0578 .

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.705 on 120 degrees of freedom

Multiple R-squared:  0.6088,           Adjusted R-squared:  0.6023

F-statistic: 93.37 on 2 and 120 DF,  p-value: < 2.2e-16

> # if we use alpha = 0.05, then VALUE exceeds our cutoff, so we'll remove that too.

> model3 = lm(VOLUME ~ SHARE)

> summary(model3)

Call:

lm(formula = VOLUME ~ SHARE)

Residuals:

    Min      1Q  Median      3Q     Max

-22.153  -3.778  -1.226   3.031  27.881

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.662394   0.792936   8.402 9.88e-14 \*\*\*

SHARE       0.071363   0.005332  13.384  < 2e-16 \*\*\*

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.779 on 121 degrees of freedom

Multiple R-squared:  0.5968,           Adjusted R-squared:  0.5935

F-statistic: 179.1 on 1 and 121 DF,  p-value: < 2.2e-16

> #Now we see that share is significant, so we would stop. THe loss that you have in r2 is about 60% in each. Havne't lost a lot in explanitory power. So if you had some personal choice, if you really wanted to include value in the model, just pick alpha to be 0.10 so it would be significant.

> # Here stepwise and backwards chose different models, but they will often choose the same one.

>

> liquid1 = data.frame(AVGT, NTRAN, PRICE, SHARE, VALUE, DEBEQ, VOLUME)

> pairs(liquid1, upper.panel=NULL)

>

> # raw residuals are below

> residuals(model1)

          1           2           3           4           5           6

  3.8423812  -3.3614749  -6.9201979  -4.3595637   1.4123279  -7.2709527

          7           8           9          10          11          12

 -0.2939946  -6.8611451   5.7651224   6.2439693   9.8631219   1.6878731

         13          14          15          16          17          18

 12.3714398  -4.1090876  -3.9132002  -0.6615502  -0.8594588  -1.0911598

         19          20          21          22          23          24

  0.8709216  -1.0050691   5.3639545  -5.9260785   4.2255715   2.7690669

         25          26          27          28          29          30

 -0.4969370  -4.1599465  -5.3212745  -2.0642073  -0.5476797  -2.8320358

         31          32          33          34          35          36

  7.3829295  -5.0704648   0.5736178   4.3036987  -6.1596472   8.9386865

         37          38          39          40          41          42

 28.2295406  -4.9967245   7.9049777  -3.0035114  -2.2800158  11.7146468

         43          44          45          46          47          48

 -2.6119224  -2.4800348 -10.5728446  -4.2631327  -5.7986429   6.8703743

         49          50          51          52          53          54

  1.4178419   2.7211179   5.8058681   5.7249694 -20.7076969  -4.3063679

         55          56          57          58          59          60

  2.2906868  -3.5399362  -5.0360661   3.2719911   9.9137928   2.8795743

         61          62          63          64          65          66

 -6.5834226  -4.9922705   6.0008537  18.4059021   2.7183615  10.1383016

         67          68          69          70          71          72

 -2.2703476  -0.8525352  -4.6026237  -4.4909554  -5.4528692  -2.4038518

         73          74          75          76          77          78

 -2.0006556  -5.4112683  -5.7120791  -3.1855565   5.2588586  -1.1686512

         79          80          81          82          83          84

  1.0448060   2.9435745  -3.9961649   7.7523754  -1.3785135  -2.7337243

         85          86          87          88          89          90

 21.3870984  -0.1277991   0.4566533   3.3627457  -3.9079901   0.8687484

         91          92          93          94          95          96

 -4.1972059   1.5926042  -3.0466414   1.1401375  -3.7767303  -5.4904436

         97          98          99         100         101         102

  0.2720958  -1.1918046   0.3599378   0.3800121   4.1102770  -6.8413731

        103         104         105         106         107         108

 12.4724174  -6.4160716  -6.8965048  -3.6789801   7.3904512  -1.9417017

        109         110         111         112         113         114

 -4.7027410  -5.3705189   8.3790712  -0.5643730  10.1607048  -3.5735810

        115         116         117         118         119         120

  1.2633487  -5.2452480  -0.1519795  -2.7579383  -3.4415784  -2.7607495

        121         122         123

 -3.6985129 -19.4123096   7.0908865

> # Standardized residuals are below and then put in order

> rstdandard = rstandard(model1)

> rstandard = rstandard(model1)

> rstandard[order(rstandard)]

        122          53          45           6         105           3

-3.54676310 -3.14586376 -1.64788736 -1.09627847 -1.05365175 -1.04390441

          8         102          61         104          35          22

-1.03736974 -1.03140944 -0.98573369 -0.96885735 -0.92298032 -0.89105514

         47          75          71          96          74         110

-0.87769161 -0.86329189 -0.86284826 -0.82020931 -0.81152955 -0.80744595

         27         116          32          57          38          62

-0.80313404 -0.78570203 -0.77041769 -0.75751353 -0.75042005 -0.74692143

        109          69          70           4          54          46

-0.70715297 -0.69366257 -0.67330955 -0.65450240 -0.64564946 -0.63937522

         91          26          14          81          15          89

-0.63288165 -0.62394667 -0.61512277 -0.60200281 -0.58546706 -0.58462006

         95         121         106         114          56         119

-0.56484034 -0.55343731 -0.55125177 -0.54094739 -0.53003324 -0.51517253

          2          76          93          40          30         118

-0.50446509 -0.48315322 -0.45634892 -0.45107653 -0.43384966 -0.41536402

        120          84          43          44          72          41

-0.41282855 -0.41026465 -0.39203937 -0.37494952 -0.36071976 -0.34197201

         67          28          73         108          83          78

-0.33970532 -0.30917037 -0.29939719 -0.29078377 -0.20608651 -0.18035419

         98          18          20          17          68          16

-0.17990024 -0.16314362 -0.15245893 -0.12865093 -0.12800632 -0.09911285

        112          29          25           7         117          86

-0.09211513 -0.08225064 -0.07436267 -0.04393585 -0.02278210 -0.01910017

         97          99         100          87          33          90

 0.04069393  0.05376894  0.05895020  0.06834189  0.08628492  0.12990611

         19          79          94         115           5          49

 0.13081276  0.15829590  0.17094555  0.18881039  0.21117400  0.21202216

         92          12          55          65          50          24

 0.23794261  0.25386573  0.34473174  0.40762550  0.40773555  0.41384557

         80          58          88           1         101          23

 0.43985547  0.48983943  0.50394548  0.57398325  0.61491229  0.63283548

         34          21          77          52           9          51

 0.64697343  0.80216824  0.83491808  0.85648764  0.86639232  0.86826470

         63          10          48         123          31         107

 0.89931201  0.93524298  1.05715926  1.05923762  1.10432269  1.10465554

         82          39          60         111          36          11

 1.15962268  1.18186613  1.21451250  1.25228163  1.34901247  1.47776804

         59          66         113          42          13         103

 1.50058397  1.51549195  1.52015032  1.75086027  1.86374971  1.89757181

         64          85          37

 2.75551066  3.21646312  4.22749718

> # outliers that exceed | 2 | are 64, 85, 37, 122, 53 observation numbers

> rstudent = rstudent(model1)

> rstudent[order(rstudent)]

        122          53          45           6         105           3

-3.73474192 -3.27160972 -1.65999841 -1.09721721 -1.05414402 -1.04430157

          8         102          61         104          35          22

-1.03770457 -1.03168841 -0.98561538 -0.96860572 -0.92240163 -0.89027830

         47          75          71          96          74         110

-0.87683876 -0.86236160 -0.86191565 -0.81907430 -0.81035806 -0.80625783

         27         116          32          57          38          62

-0.80192874 -0.78443111 -0.76909424 -0.75614929 -0.74903476 -0.74552612

        109          69          70           4          54          46

-0.70565970 -0.69214260 -0.67175534 -0.65292284 -0.64406000 -0.63777953

         91          26          14          81          15          89

-0.63128017 -0.62233834 -0.61350891 -0.60038296 -0.58384340 -0.58299631

         95         121         106         114          56         119

-0.56321757 -0.55181766 -0.54963292 -0.53933323 -0.52842563 -0.51357640

          2          76          93          40          30         118

-0.50287901 -0.48159147 -0.45482560 -0.44956176 -0.43236499 -0.41391527

        120          84          43          44          72          41

-0.41138501 -0.40882644 -0.39064102 -0.37359152 -0.35939747 -0.34069957

         67          28          73         108          83          78

-0.33843911 -0.30799231 -0.29824892 -0.28966234 -0.20525540 -0.17961935

         98          18          20          17          68          16

-0.17916712 -0.16247487 -0.15183183 -0.12811815 -0.12747612 -0.09869960

        112          29          25           7         117          86

-0.09173054 -0.08190665 -0.07405129 -0.04375121 -0.02268622 -0.01901978

         97          99         100          87          33          90

 0.04052287  0.05354320  0.05870285  0.06805547  0.08592430  0.12936831

         19          79          94         115           5          49

 0.13027133  0.15764599  0.17024668  0.18804356  0.21032426  0.21116932

         92          12          55          65          50          24

 0.23699713  0.25286530  0.34345177  0.40619285  0.40630267  0.41239993

         80          58          88           1         101          23

 0.43835993  0.48826944  0.50235993  0.57235931  0.61329831  0.63123396

         34          21          77          52           9          51

 0.64538537  0.80095914  0.83384850  0.85552235  0.86547831  0.86736063

         63          10          48         123          31         107

 0.89858416  0.93474674  1.05768638  1.05978554  1.10535137  1.10568798

         82          39          60         111          36          11

 1.16132026  1.18385834  1.21696458  1.25530764  1.35372333  1.48523691

         59          66         113          42          13         103

 1.50860701  1.52388817  1.52866497  1.76638825  1.88359663  1.91883568

         64          85          37

 2.83587149  3.35193810  4.56654821

> # outliers that exceed | 2 | are 64, 85, 37, 122, 53 observation numbers. Numbers are a little differnt, but the choices are the same of which ones are outliers

> modelb = lm(VOLUME ~ PRICE + SHARE + VALUE, subset=-c(64, 85, 122, 37, 53))

> summary(modelb)

Call:

lm(formula = VOLUME ~ PRICE + SHARE + VALUE, subset = -c(64,

    85, 122, 37, 53))

Residuals:

    Min      1Q  Median      3Q     Max

-15.465  -3.125  -1.096   3.074  11.524

Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept)  5.660212   1.213632   4.664 8.50e-06 \*\*\*

PRICE       -0.005026   0.025789  -0.195    0.846

SHARE        0.076089   0.008802   8.645 3.88e-14 \*\*\*

VALUE        0.157280   0.121947   1.290    0.200

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.912 on 114 degrees of freedom

Multiple R-squared:  0.748, Adjusted R-squared:  0.7414

F-statistic: 112.8 on 3 and 114 DF,  p-value: < 2.2e-16

> leverages = hatvalues(model1)

> leverages[order(leverages)]

        123           1          99          96          92          80

0.008258723 0.008285688 0.008309569 0.008370130 0.008585244 0.008907493

          7         115          24         111          86          42

0.009111166 0.009216457 0.009226453 0.009232257 0.009247409 0.009305989

        107          66          83          39          18           5

0.009459419 0.009607680 0.009836216 0.009968895 0.010033212 0.010138859

         90         120          49          21          51          97

0.010278509 0.010311164 0.010359963 0.010482779 0.010503757 0.010607662

         95          31          82          89         101          52

0.010616514 0.010878673 0.010945303 0.011119983 0.011219393 0.011245998

        113          15          62          67         121          25

0.011313250 0.011348340 0.011377504 0.011528528 0.011671074 0.011727296

         73          87          17         119          14          58

0.011826179 0.011940056 0.012337970 0.012370535 0.012467734 0.012582427

         64          56          61          37         108          23

0.012597690 0.012882656 0.012883559 0.013209223 0.013246192 0.013326685

         28          10          93         116          16          11

0.013504711 0.013591687 0.013646179 0.013719583 0.014061668 0.014173749

        106          50          35          88          63         117

0.014313065 0.014352363 0.014376832 0.014617004 0.014652540 0.015157690

         70          54          94          65          74          46

0.015462985 0.015506623 0.015574359 0.015817011 0.016048785 0.016146985

         41          26          72           2          84          43

0.016263672 0.016295493 0.017212496 0.017389004 0.017425851 0.017697417

          4          68          29          38          40          19

0.018144247 0.018374287 0.018797153 0.018825904 0.018834622 0.019060992

          9          34         110          22         109          85

0.020122020 0.020746916 0.020984208 0.021164412 0.021277662 0.021567889

         12          57          33          55         118          81

0.021737184 0.021893452 0.021953266 0.022868562 0.024346862 0.024844499

         13          69         102           6          91           3

0.024899397 0.025686932 0.026337459 0.026524593 0.026671024 0.027477450

         36          27          98         104          75          44

0.028372951 0.028507689 0.028750606 0.029483679 0.031148493 0.031824524

          8          47          59         114          79          76

0.031922399 0.034053782 0.034075504 0.034213090 0.035913393 0.037978009

         20          53          32         103         105          30

0.038232279 0.041111986 0.041422652 0.043931637 0.051913379 0.057018720

         48          78         100          45          71          77

0.065317580 0.070816532 0.080380340 0.089011776 0.116176737 0.122030615

        112         122          60

0.169280041 0.337060086 0.875595321

> dim(liquid)

[1] 123   9

> # n = 123 and k=3

> # high leverage cutoff = 3 (k+1)/n = 3(3+1)/123 = 0.09756

> # high leverage point are ovservations 71, 77, 112, 122, 60

> summary(modelb)

Call:

lm(formula = VOLUME ~ PRICE + SHARE + VALUE, subset = -c(64,

    85, 122, 37, 53))

Residuals:

    Min      1Q  Median      3Q     Max

-15.465  -3.125  -1.096   3.074  11.524

Coefficients:

             Estimate Std. Error t value Pr(>|t|)

(Intercept)  5.660212   1.213632   4.664 8.50e-06 \*\*\*

PRICE       -0.005026   0.025789  -0.195    0.846

SHARE        0.076089   0.008802   8.645 3.88e-14 \*\*\*

VALUE        0.157280   0.121947   1.290    0.200

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.912 on 114 degrees of freedom

Multiple R-squared:  0.748, Adjusted R-squared:  0.7414

F-statistic: 112.8 on 3 and 114 DF,  p-value: < 2.2e-16

> model1c=lm(VOLUME ~ PRICE + SHARE + VALUE, subset=-c(64,85,122,37,53,71,77,112,60))

> summary(model1c)

Call:

lm(formula = VOLUME ~ PRICE + SHARE + VALUE, subset = -c(64,

    85, 122, 37, 53, 71, 77, 112, 60))

Residuals:

    Min      1Q  Median      3Q     Max

-13.765  -3.296  -1.352   3.066  11.199

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  4.58802    1.52978   2.999  0.00335 \*\*

PRICE        0.02550    0.03622   0.704  0.48282

SHARE        0.08686    0.01511   5.747 8.21e-08 \*\*\*

VALUE       -0.12423    0.32464  -0.383  0.70270

---

Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.954 on 110 degrees of freedom

Multiple R-squared:  0.6356,           Adjusted R-squared:  0.6257

F-statistic: 63.96 on 3 and 110 DF,  p-value: < 2.2e-16

> hist(rstandard)

> plot(residuals(model1) ~ fitted.values(model1), main="Residuals vs. Fitted Values")

> #Want this to be a random scatter. If not, the indication that there is something in the dataset that shouldn't be there.

> #This graph is for model1, before we removed the points.

> # Want to see homoscedasticity. Do we think this has equal spread?

> # There seems to be random scatter and no obvious pattern here, which is good, this means that there is not some obvious X we are not accounting for.

> # There seems to also be roughly equal variance or spread throughout the plot. This is also good because homoscedasticity is an assumption of regression. This is often hard to judge.