Regression Analysis: Case Study 1

Dr. Kempthorne

September 23, 2013

Contents

1 Linear Regression Models for Asset Pricing

1.1 CAPM Theory

Sharpe (1964) and Lintner (1965) developed the Capital Asset Pricing Model for a market in which investors have the same expectations, hold portfolios of risky assets that are mean-variance efficient, and can borrow and lend money freely at the same risk-free rate. In such a market, the expected return of asset j is

$$\begin{aligned} E[R_j] &= R_{riskfree} + \beta_j (E[R_{Market}] - R_{riskfree}) \\ \beta_j &= Cov[R_j, R_{Market}]/Var[R_{Market}] \end{aligned}$$

where R_{Market} is the return on the market portfolio and $R_{riskfree}$ is the return on the risk-free asset.

Consider fitting the simple linear regression model of a stock's daily excess return on the market-portfolio daily excess return, using the S&P 500 Index as the proxy for the market return and the 3-month Treasury constant maturity rate as the risk-free rate. The linear model is given by:

$$R^*_{j,t} = \alpha_j + \beta_j R^*_{Market,t} + \epsilon_{j,t},\ t = 1, 2, \ldots$$

where $\epsilon_{j,t}$ are white noise: $WN(0, \sigma^2)$

Under the assumptions of the CAPM, the regression parameters (α_j, β_j) are such that β_j is the same as in the CAPM model, and α_j is zero.

1.2 Historical Financial Data

Executing the R-script "fm_casestudy_1_0.r" creates the time-series matrix $casestudy1.data0.00$ which is available in the R-workspace "casestudy_1_0.Rdata".

```
> library("zoo")
> load("casestudy_1_0.RData")
> dim(casestudy1.data0.0)

[1] 3373   12

> names(casestudy1.data0.00)

 [1] "BAC"        "GE"         "JDSU"       "XOM"        "SP500"
 [6] "DGS3MO"     "DGS1"       "DGS5"       "DGS10"      "DAAA"
[11] "DBAA"       "DCOILWTICO"

> head(casestudy1.data0.00)

                BAC       GE   JDSU      XOM   SP500 DGS3MO DGS1 DGS5 DGS10
2000-01-03 15.79588 33.39834 752.00 28.83212 1455.22   5.48 6.09 6.50  6.58
2000-01-04 14.85673 32.06240 684.52 28.27985 1399.42   5.43 6.00 6.40  6.49
2000-01-05 15.01978 32.00674 633.00 29.82252 1402.11   5.44 6.05 6.51  6.62
2000-01-06 16.30458 32.43424 599.00 31.36519 1403.45   5.41 6.03 6.46  6.57
```

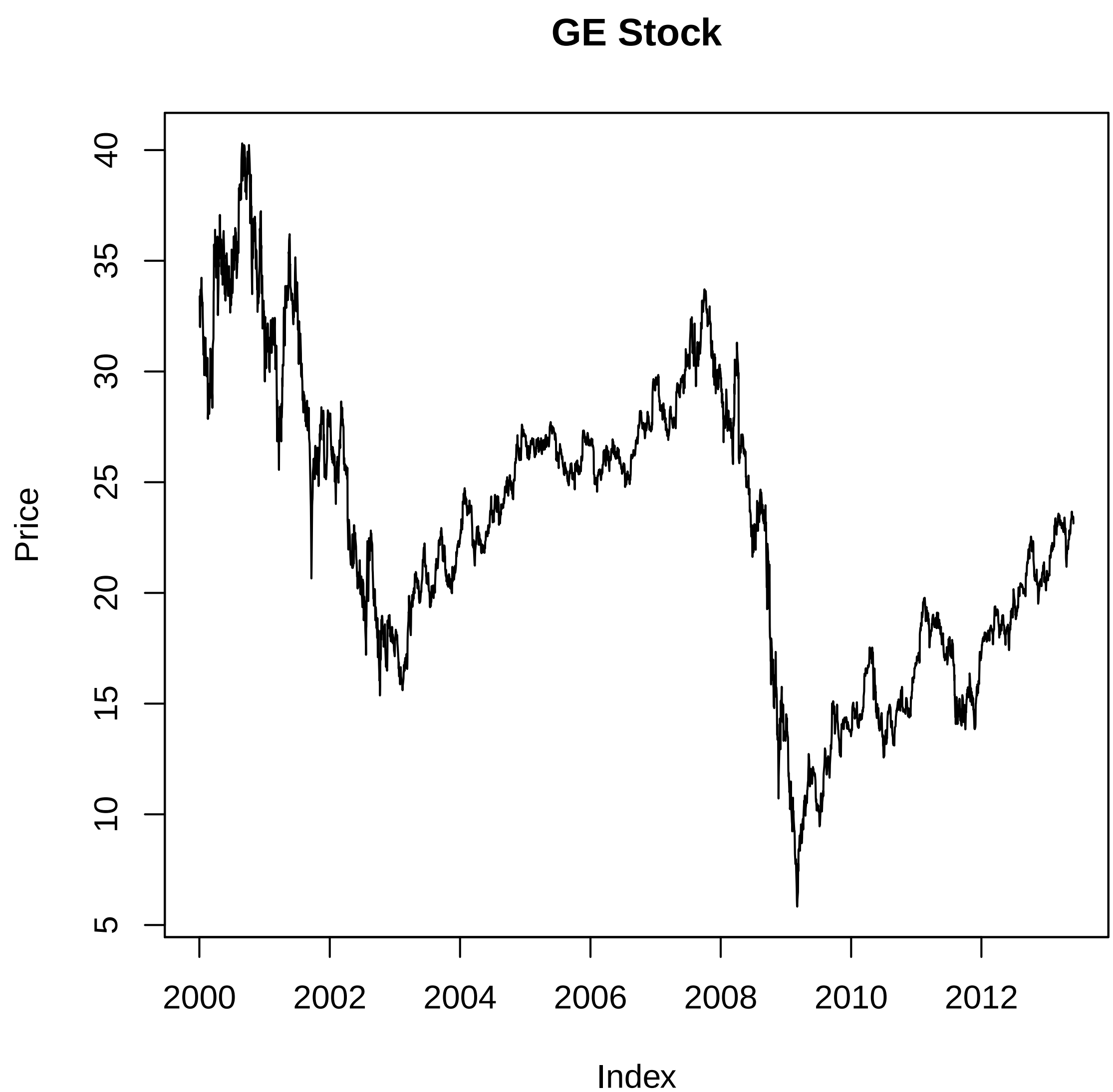
```
2000-01-07 15.87740 33.69002 719.76 31.27315 1441.47   5.38 6.00 6.42  6.52
2000-01-10 15.32631 33.67666 801.52 30.83501 1457.60   5.42 6.07 6.49  6.57
           DAAA DBAA DCOILWTICO
2000-01-03 7.75 8.27         NA
2000-01-04 7.69 8.21      25.56
2000-01-05 7.78 8.29      24.65
2000-01-06 7.72 8.24      24.79
2000-01-07 7.69 8.22      24.79
2000-01-10 7.72 8.27      24.71
```

```
> tail(casestudy1.data0.00)
```

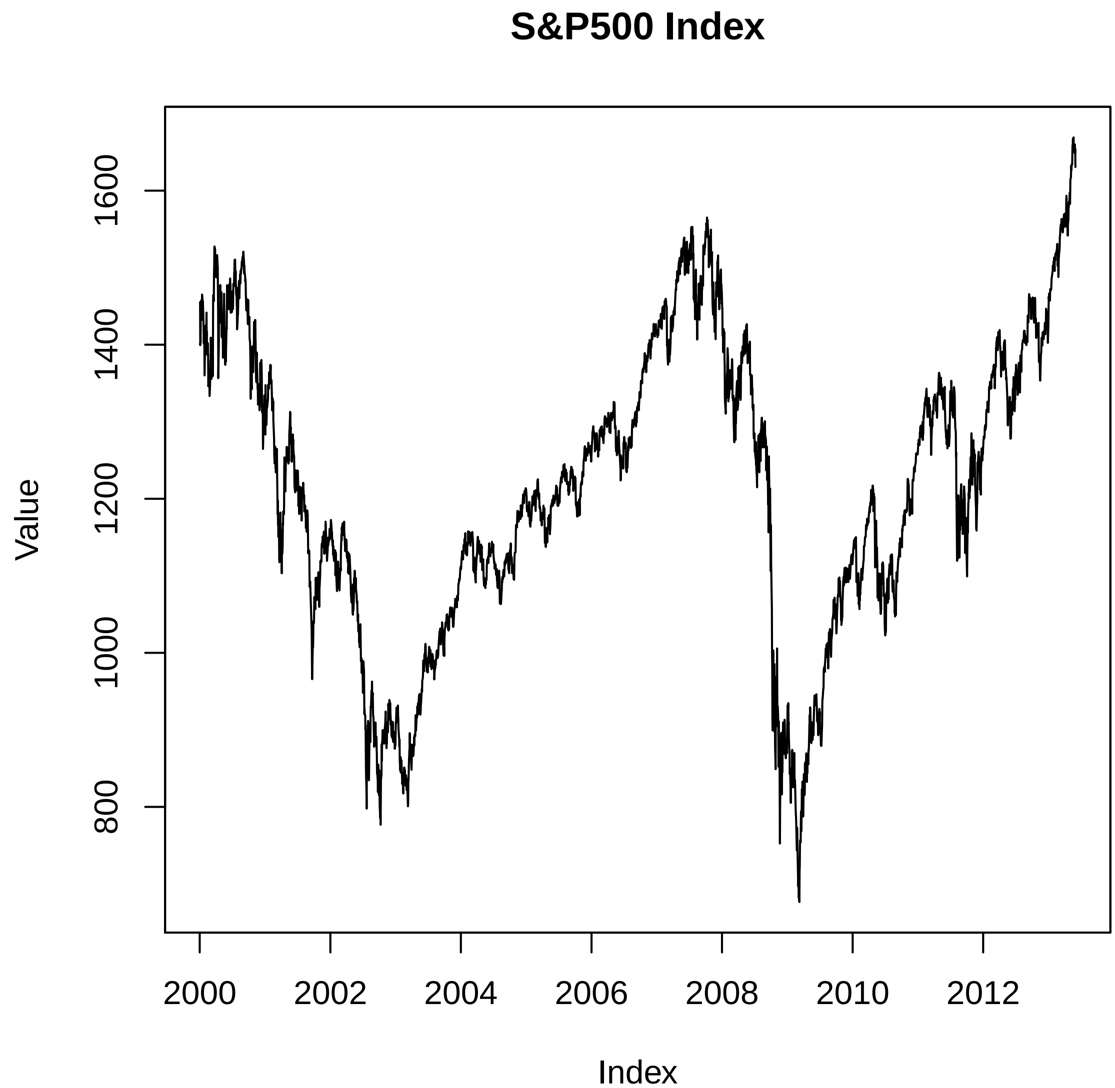
```
                BAC       GE  JDSU   XOM   SP500 DGS3MO DGS1 DGS5 DGS10 DAAA
2013-05-23 13.20011 23.47254 13.17 91.79 1650.51   0.05 0.12 0.91  2.02 3.97
2013-05-24 13.23009 23.34357 13.07 91.53 1649.60   0.04 0.12 0.90  2.01 3.94
2013-05-28 13.34001 23.41301 13.37 92.38 1660.06   0.05 0.13 1.02  2.15 4.06
2013-05-29 13.46991 23.45269 13.56 92.08 1648.36   0.05 0.14 1.02  2.13 4.04
2013-05-30 13.81965 23.41301 13.73 92.09 1654.41   0.04 0.13 1.01  2.13 4.06
2013-05-31 13.64978 23.13523 13.62 90.47 1630.74   0.04 0.14 1.05  2.16 4.09
           DBAA DCOILWTICO
2013-05-23 4.79      94.12
2013-05-24 4.76      93.84
2013-05-28 4.88      94.65
2013-05-29 4.88      93.13
2013-05-30 4.90      93.57
2013-05-31 4.95      91.93
```

We first plot the raw data for the stock GE, the market-portfolio index $SP500$, and the risk-free interest rate.

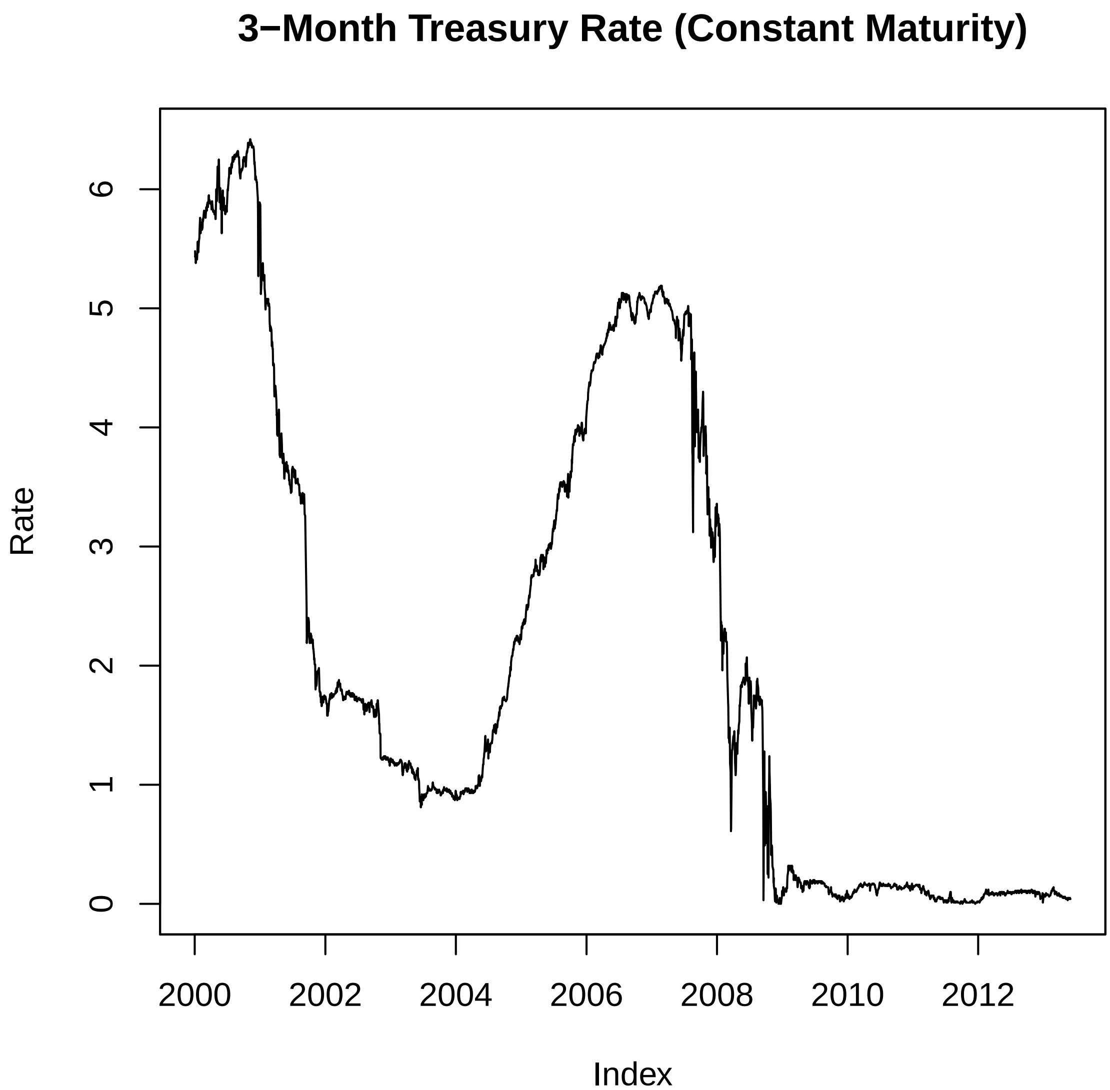
```
> library ("graphics")
> library("quantmod")
> plot(casestudy1.data0.00[,"GE"],ylab="Price",main="GE Stock")
```



```
> plot(casestudy1.data0.00[,"SP500"], ylab="Value",main="S&P500 Index")
```



```
> plot(casestudy1.data0.00[,"DGS3MO"], ylab="Rate" ,
+      main="3-Month Treasury Rate (Constant Maturity)")
```



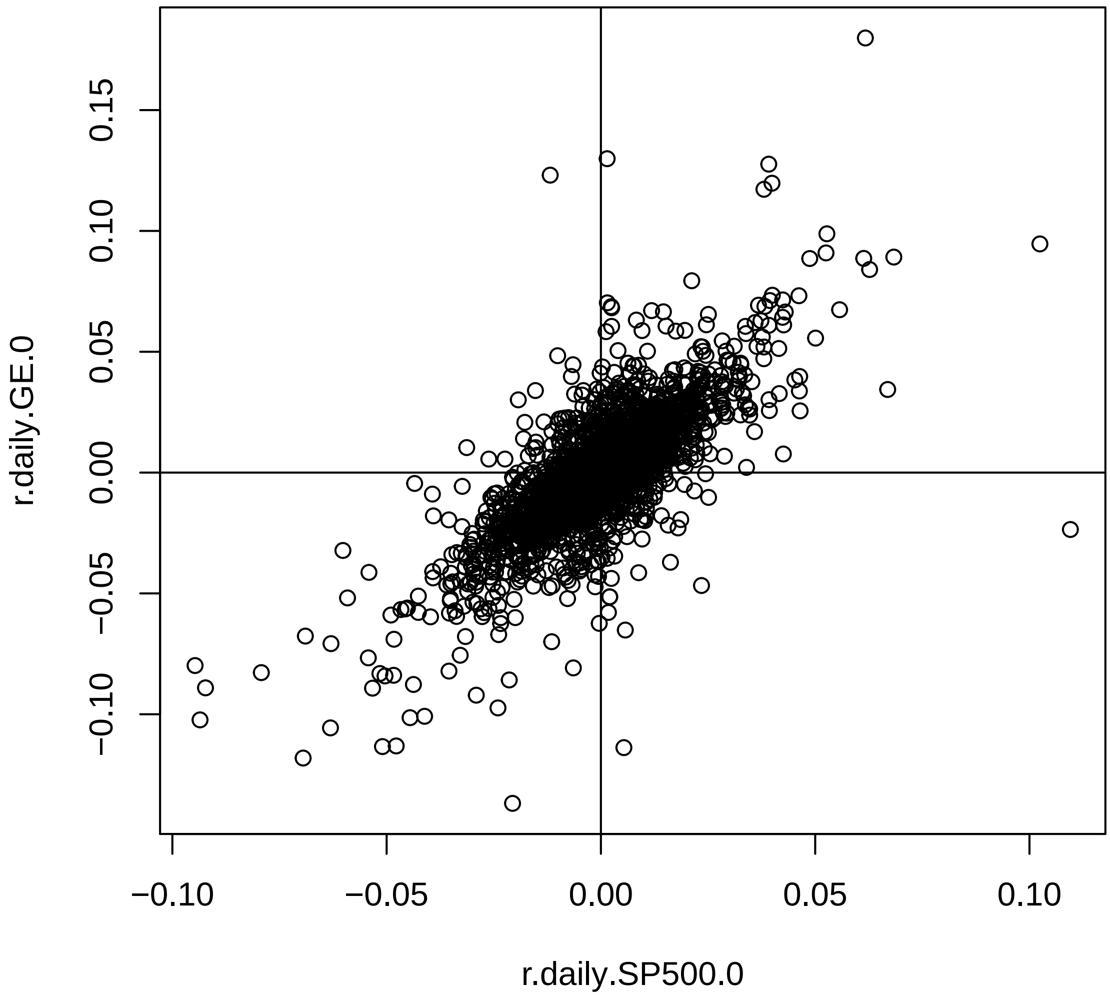
Now we construct the variables with the log daily returns of GE and the SP500 index as well as the risk-free asset returns

```
> # Compute daily log returns of GE stock
> r.daily.GE<-zoo( x=as.matrix(diff(log(casestudy1.data0.00[,"GE"]))),
```

```
+                  order.by=time(casestudy1.data0.00)[-1])
> dimnames(r.daily.GE)[[2]]<-"r.daily.GE"
> dim(r.daily.GE)

[1] 3372    1

> head(r.daily.GE)

              r.daily.GE
2000-01-04 -0.0408219945
2000-01-05 -0.0017376199
2000-01-06  0.0132681098
2000-01-07  0.0379869230
2000-01-10 -0.0003966156
2000-01-11  0.0016515280

> # Compute daily log returns of the SP500 index
> r.daily.SP500<-zoo( x=as.matrix(diff(log(casestudy1.data0.00[,"SP500"]))),
+                     order.by=time(casestudy1.data0.00)[-1])
> dimnames(r.daily.SP500)[[2]]<-"r.daily.SP500"
> dim(r.daily.SP500)

[1] 3372    1

> head(r.daily.SP500)

           r.daily.SP500
2000-01-04 -0.0390992269
2000-01-05  0.0019203798
2000-01-06  0.0009552461
2000-01-07  0.0267299353
2000-01-10  0.0111278213
2000-01-11 -0.0131486343

> # Compute daily return of the risk-free asset
> #     accounting for the number of days between successive closing prices
> #     apply annual interest rate using 360 days/year (standard on 360-day yearsince the pr
>
> r.daily.riskfree<-log(1 + .01*coredata(casestudy1.data0.00[-1,"DGS3MO"]) *
+                         diff(as.numeric(time(casestudy1.data0.00)))/360)
> dimnames(r.daily.riskfree)[[2]]<-"r.daily.riskfree"
> # Compute excess returns (over riskfree rate)
>
> r.daily.GE.0<-r.daily.GE - r.daily.riskfree
>   dimnames(r.daily.GE.0)[[2]]<-"r.daily.GE.0"
> r.daily.SP500.0<-r.daily.SP500 - r.daily.riskfree
>   dimnames(r.daily.SP500.0)[[2]]<-"r.daily.SP500.0"
```

```
> # Merge all the time series together,
> #     and display first and last sets of rows
> r.daily.data0<-merge(r.daily.GE, r.daily.SP500, r.daily.riskfree,
+                      r.daily.GE.0, r.daily.SP500.0)
> head(r.daily.data0)

             r.daily.GE r.daily.SP500 r.daily.riskfree  r.daily.GE.0
2000-01-04 -0.0408219945 -0.0390992269     0.0001508220 -0.0409728165
2000-01-05 -0.0017376199  0.0019203798     0.0001510997 -0.0018887196
2000-01-06  0.0132681098  0.0009552461     0.0001502665  0.0131178433
2000-01-07  0.0379869230  0.0267299353     0.0001494333  0.0378374897
2000-01-10 -0.0003966156  0.0111278213     0.0004515647 -0.0008481802
2000-01-11  0.0016515280 -0.0131486343     0.0001508220  0.0015007061
           r.daily.SP500.0
2000-01-04   -0.0392500488
2000-01-05    0.0017692801
2000-01-06    0.0008049796
2000-01-07    0.0265805020
2000-01-10    0.0106762566
2000-01-11   -0.0132994562

> tail(r.daily.data0)

             r.daily.GE r.daily.SP500 r.daily.riskfree r.daily.GE.0
2013-05-23 -0.008417558 -0.0029281358     1.388888e-06 -0.008418947
2013-05-24 -0.005509656 -0.0005514968     1.111110e-06 -0.005510767
2013-05-28  0.002970509  0.0063209120     5.555540e-06  0.002964954
2013-05-29  0.001693481 -0.0070728921     1.388888e-06  0.001692092
2013-05-30 -0.001693481  0.0036635956     1.111110e-06 -0.001694592
2013-05-31 -0.011935351 -0.0144105503     1.111110e-06 -0.011936462
           r.daily.SP500.0
2013-05-23   -0.0029295247
2013-05-24   -0.0005526079
2013-05-28    0.0063153565
2013-05-29   -0.0070742809
2013-05-30    0.0036624845
2013-05-31   -0.0144116614
```

Now we plot the excess returns of GE vs those of the SP500:



1.3 Fitting the Linear Regression for CAPM

The linear regression model is fit using the R-function $lm()$:

```
>   lmfit0<-lm(r.daily.GE.0 ~ r.daily.SP500.0, data=r.daily.data0)
>   names(lmfit0) #element names of list object lmfit0
```

```
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

>   summary.lm(lmfit0) #function summarizing objects created by lm()

Call:
lm(formula = r.daily.GE.0 ~ r.daily.SP500.0, data = r.daily.data0)

Residuals:
      Min        1Q    Median        3Q       Max
-0.153166 -0.005605 -0.000334  0.005560  0.137232

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.0001334  0.0002376  -0.561    0.575
r.daily.SP500.0  1.1843613  0.0177920  66.567   <2e-16

Residual standard error: 0.0138 on 3370 degrees of freedom
Multiple R-squared: 0.568,       Adjusted R-squared: 0.5679
F-statistic:  4431 on 1 and 3370 DF,  p-value: < 2.2e-16

>
```

Note that the t-statistic for the intercept α_{GE} is not significant (-0.5613).

1.4 Regression Diagnostics

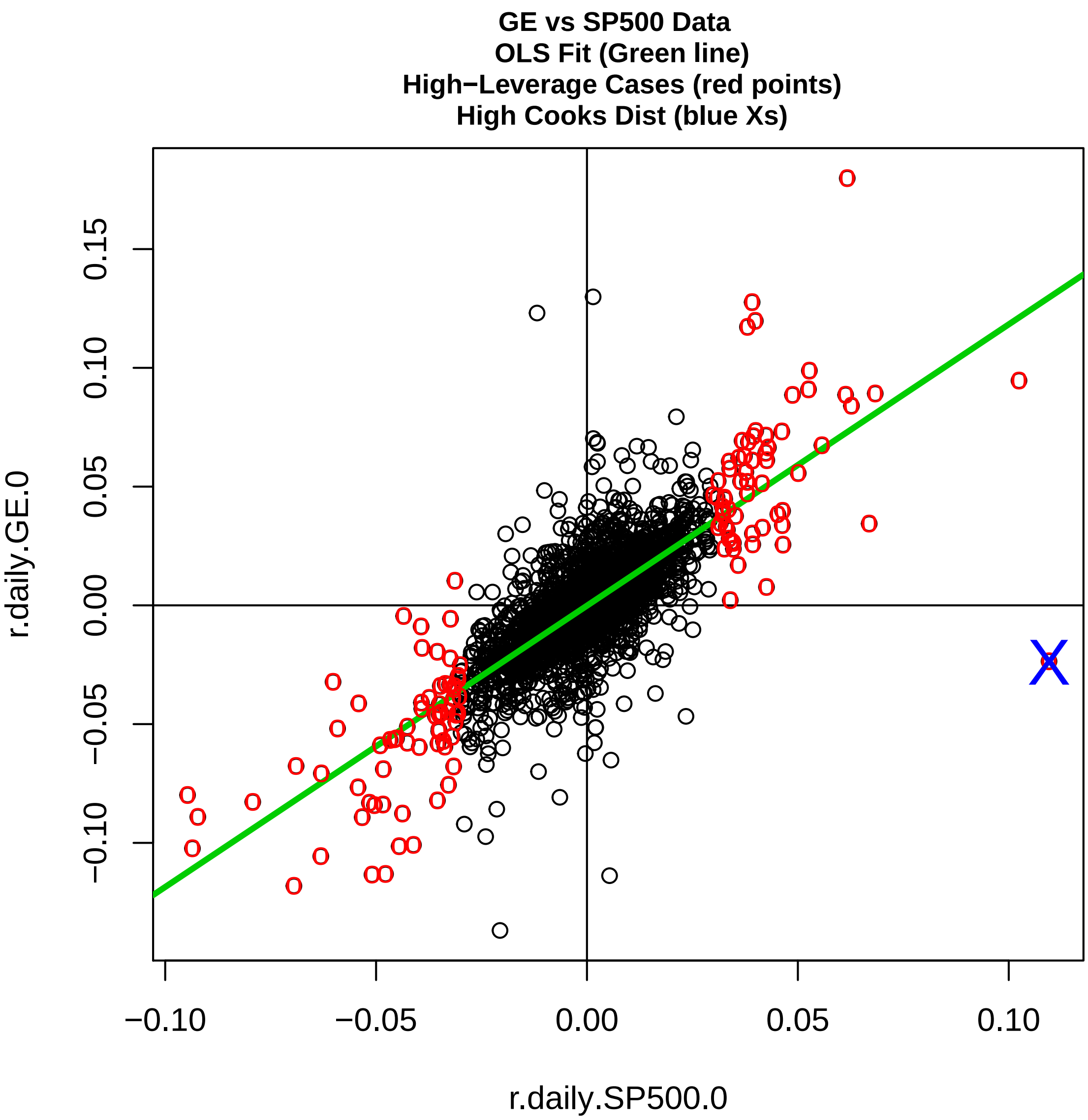
Some useful R functions

- *anova.lm*(): conduct an Analysis of Variance for the linear regression model, detailing the computation of the F-statistic for no regression structure.
- *influence.measures*(): compute regression diagnostics evaluating case influence for the linear regression model; includes 'hat' matirx, case-deletion statistics for the regression coefficients and for the residual standard deviation.

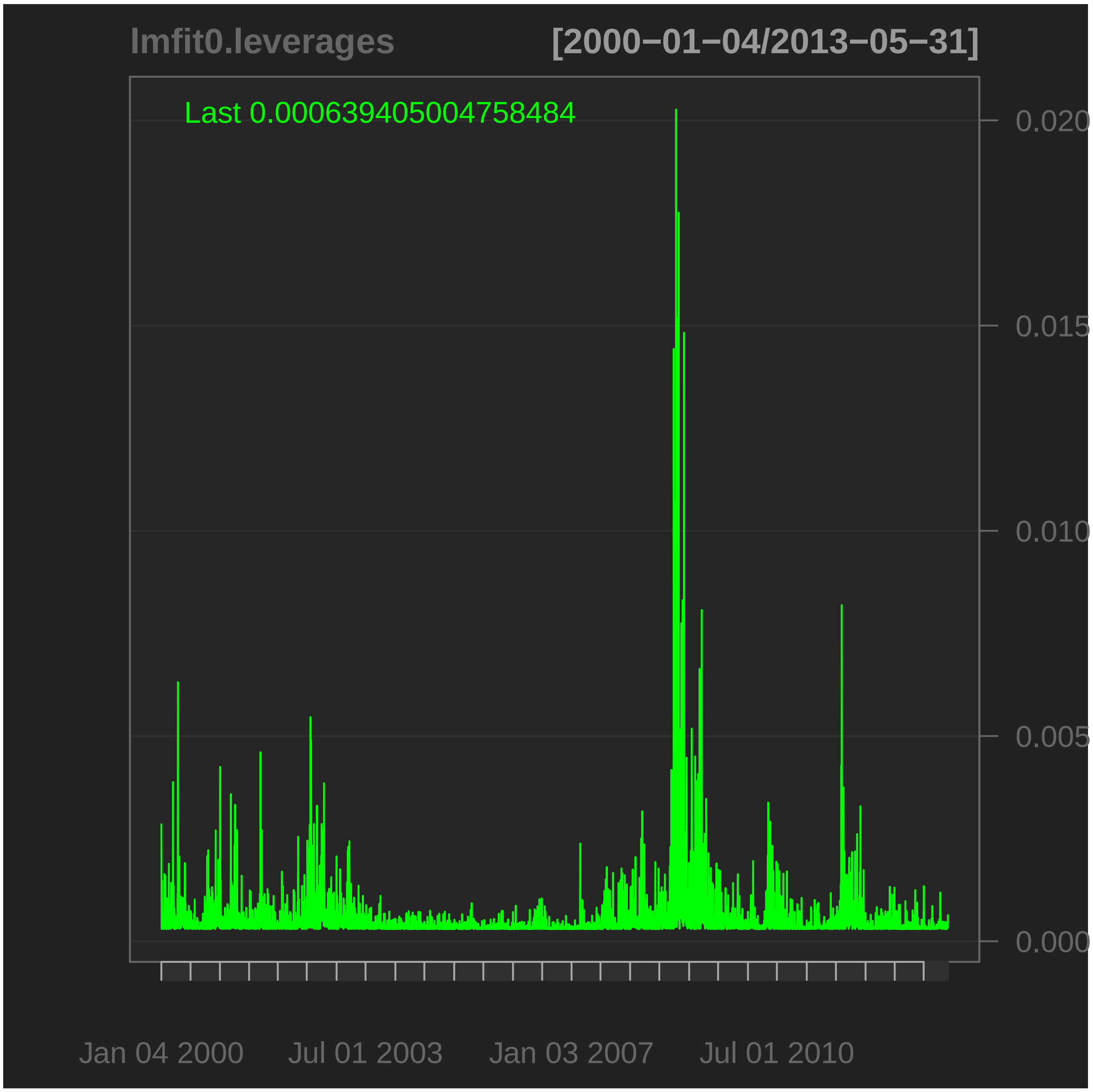
```
> # Compute influence measures (case-deletion statistics)
> lmfit0.inflm<-influence.measures(lmfit0)
> names(lmfit0.inflm)

[1] "infmat" "is.inf" "call"

> dim(lmfit0.inflm$infmat)

[1] 3372    6
```

```
> head(lmfit0.inflm$infmat)

                dfb.1_      dfb.r..S        dffit    cov.r       cook.d
2000-01-04  0.006987967 -0.0207373156  0.021908094 1.003354 2.400416e-04
2000-01-05 -0.004808670 -0.0006547183 -0.004850631 1.000850 1.176753e-05
2000-01-06  0.015354314  0.0009828679  0.015382160 1.000420 1.183126e-04
2000-01-07  0.008170676  0.0161694450  0.018089276 1.001941 1.636488e-04
2000-01-10 -0.016729492 -0.0133945658 -0.021391856 1.000525 2.288100e-04
2000-01-11  0.021629043 -0.0215352003  0.030579350 1.000239 4.674667e-04
                   hat
2000-01-04 0.0028508517
2000-01-05 0.0003020630
2000-01-06 0.0002977757
2000-01-07 0.0014754368
2000-01-10 0.0004878168
2000-01-11 0.0005883587

> head(lmfit0.inflm$is.inf)

           dfb.1_ dfb.r..S dffit cov.r cook.d   hat
2000-01-04  FALSE    FALSE FALSE  TRUE  FALSE  TRUE
2000-01-05  FALSE    FALSE FALSE FALSE  FALSE FALSE
2000-01-06  FALSE    FALSE FALSE FALSE  FALSE FALSE
2000-01-07  FALSE    FALSE FALSE  TRUE  FALSE FALSE
2000-01-10  FALSE    FALSE FALSE FALSE  FALSE FALSE
2000-01-11  FALSE    FALSE FALSE FALSE  FALSE FALSE

> # Table counts of influential/non-influential cases
> # as measured by the hat/leverage statistic.
> table(lmfit0.inflm$is.inf[,"hat"])

FALSE  TRUE
 3243   129
```

```
> # Re-Plot data adding
> #     fitted regression line
> #     selective highlighting of influential cases
>
> plot(r.daily.SP500.0, r.daily.GE.0,
+      main="GE vs SP500 Data \n OLS Fit (Green line)\n High-Leverage Cases (red points)\n H
> abline(h=0,v=0)
> abline(lmfit0, col=3, lwd=3)
> # Plot cases with high leverage as red (col=2) "o"s
> index.inf.hat<-which(lmfit0.inflm$is.inf[,"hat"]==TRUE)
> points(r.daily.SP500.0[index.inf.hat], r.daily.GE.0[index.inf.hat],
+         col=2, pch="o")
> # Plot cases with high cooks distance as big (cex=2) blue (col=4) "X"s
> index.inf.cook.d<-which(lmfit0.inflm$is.inf[,"cook.d"]==TRUE)
> points(r.daily.SP500.0[index.inf.cook.d], r.daily.GE.0[index.inf.cook.d],
+        col=4, pch="X", cex=2.)
```

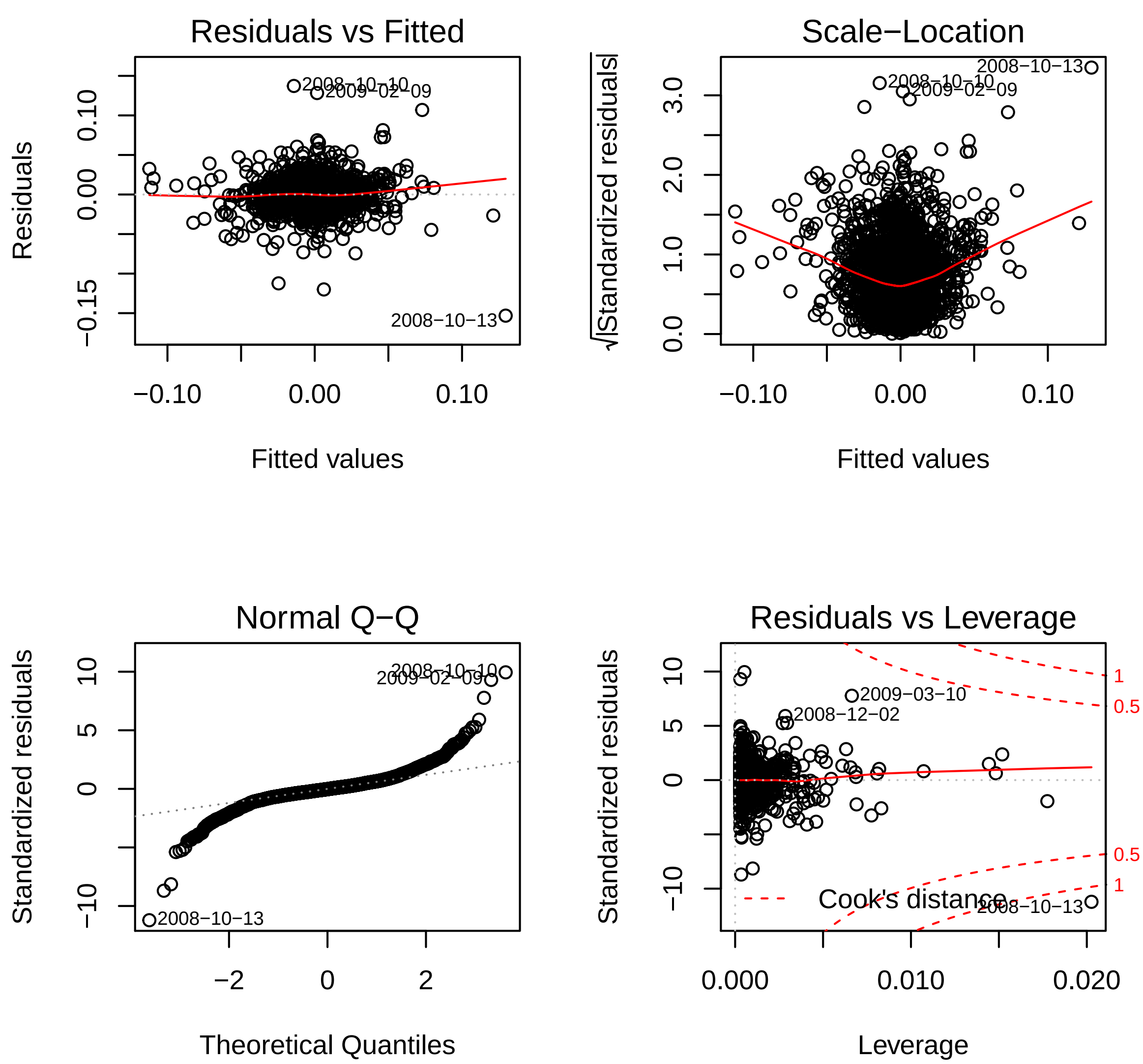

GE vs SP500 Data
OLS Fit (Green line)
High−Leverage Cases (red points)
High Cooks Dist (blue Xs)
r.daily.GE.0
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
0.10
r.daily.SP500.0

```
> lmfit0.leverages<-zoo(lmfit0.inflm$infmat[,"hat"], order.by=time(r.daily.SP500.0))
> chartSeries(lmfit0.leverages)
```



The R function *plot.lm*() generates a useful 2x2 display of plots for various regression diagnostic statistics:

```
> layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
> plot(lmfit0)
```



1.5 Adding Macro-economic Factors to CAPM

The CAPM relates a stock's return to that of the diversified market portfolio, proxied here by the S&P 500 Index. A stock's return can depend on macro-economic factors, such commodity prices, interest rates, economic growth (GDP).

```
> # The linear regression for the extended CAPM:
> lmfit1<-lm( r.daily.GE.0 ~ r.daily.SP500.0 + r.daily.DCOILWTICO, data=r.daily.data00)
> summary.lm(lmfit1)

Call:
lm(formula = r.daily.GE.0 ~ r.daily.SP500.0 + r.daily.DCOILWTICO,
    data = r.daily.data00)

Residuals:
      Min        1Q    Median        3Q       Max
-0.152977 -0.005567 -0.000260  0.005589  0.133583

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.0001216  0.0002373  -0.512 0.608532
r.daily.SP500.0     1.1972374  0.0181296  66.038  < 2e-16
r.daily.DCOILWTICO -0.0342538  0.0096188  -3.561 0.000374

Residual standard error: 0.01378 on 3368 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.5692,       Adjusted R-squared: 0.5689
F-statistic:  2225 on 2 and 3368 DF,  p-value: < 2.2e-16
```

The regression coefficient for the oil factor ($r.daily.DCOILWTICO$) is statistically significant and negative. Over the analysis period, price changes in GE stock are negatively related to the price changes in oil.

Consider the corresponding models for Exxon-Mobil stock, XOM

```
> # The linear regression for the simple CAPM:
> lmfit0<-lm( r.daily.XOM.0 ~ r.daily.SP500.0 , data=r.daily.data00)
> summary.lm(lmfit0)

Call:
lm(formula = r.daily.XOM.0 ~ r.daily.SP500.0, data = r.daily.data00)

Residuals:
      Min        1Q    Median        3Q       Max
-0.085289 -0.005788 -0.000009  0.006230  0.113614

Coefficients:
```

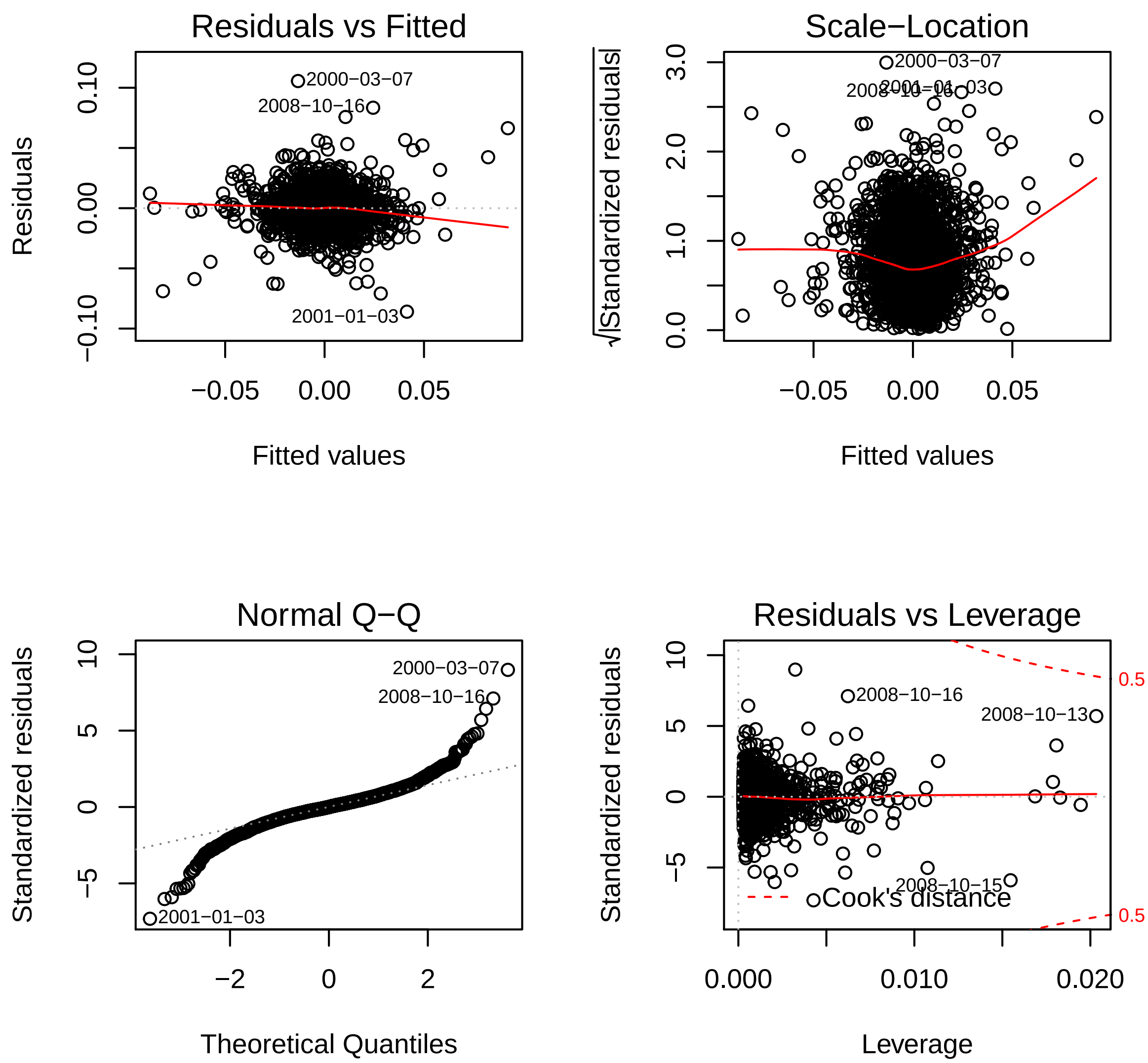
```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.0002968  0.0002105    1.41    0.159
r.daily.SP500.0 0.8299221  0.0157595   52.66   <2e-16

Residual standard error: 0.01222 on 3370 degrees of freedom
Multiple R-squared: 0.4514,        Adjusted R-squared: 0.4513
F-statistic:  2773 on 1 and 3370 DF,  p-value: < 2.2e-16

> # The linear regression for the extended CAPM:
> lmfit1<-lm( r.daily.XOM.0 ~ r.daily.SP500.0 + r.daily.DCOILWTICO.0, data=r.daily.data00)
> summary.lm(lmfit1)

Call:
lm(formula = r.daily.XOM.0 ~ r.daily.SP500.0 + r.daily.DCOILWTICO.0,
    data = r.daily.data00)

Residuals:
      Min        1Q    Median        3Q       Max
-0.085977 -0.005564  0.000010  0.005765  0.105583

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.0002520  0.0002029   1.242    0.214
r.daily.SP500.0      0.7823785  0.0155009  50.473   <2e-16
r.daily.DCOILWTICO.0 0.1324461  0.0082237  16.105   <2e-16

Residual standard error: 0.01178 on 3368 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.4906,        Adjusted R-squared: 0.4903
F-statistic:  1622 on 2 and 3368 DF,  p-value: < 2.2e-16
```

The R-squared for XOM is lower than for GE. Its relationship to the market index is less strong.

The regression coefficient for the oil factor ($r.daily.DCOILWTICO$) is statistically significant and positive.

For the extended model, we use the R function *plot.lm*() to display regression diagnostic statistics:

```
> layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
> plot(lmfit1)
```



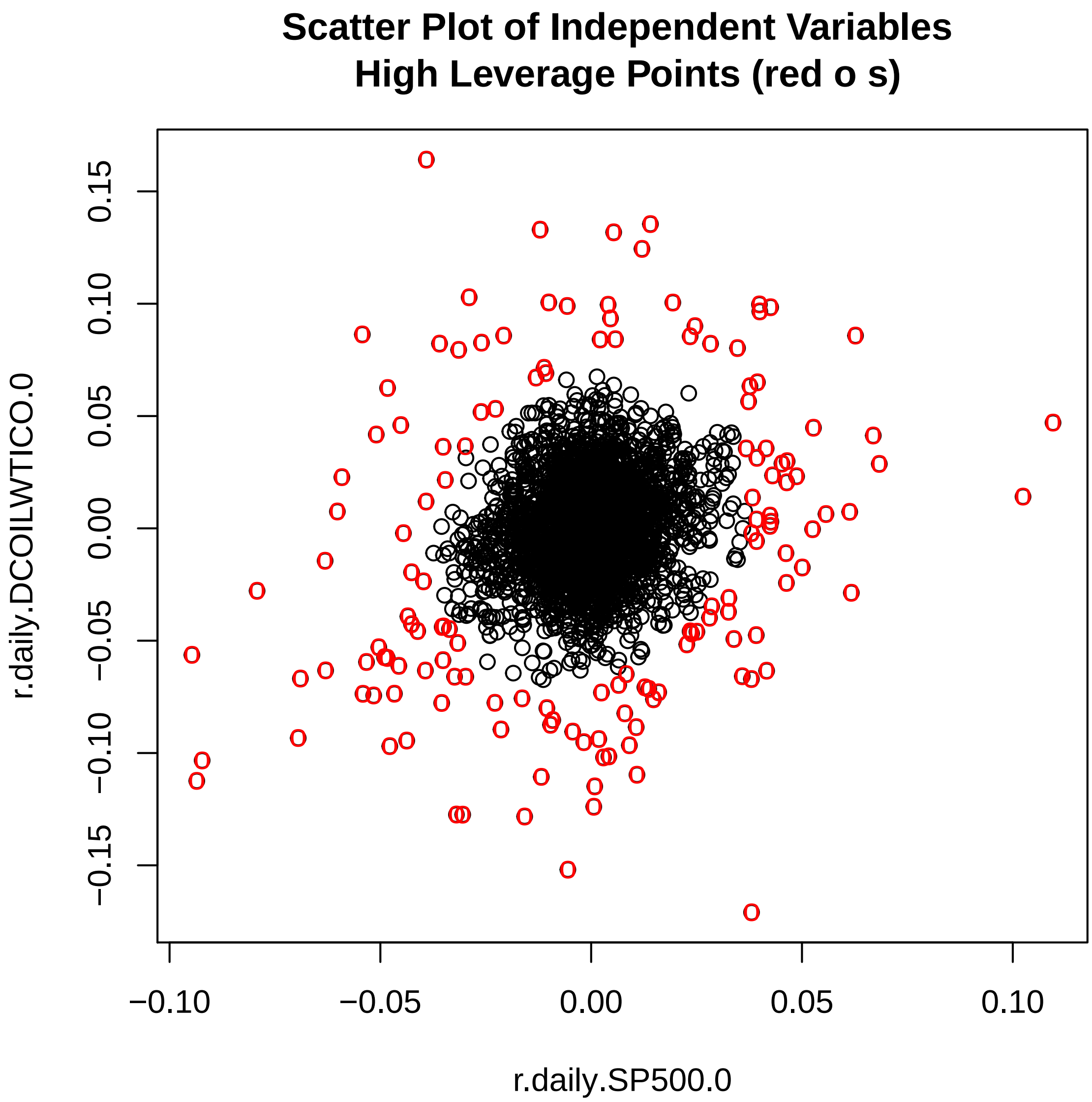
The high-leverage cases in the data are those which have high Mahalanobis distance from the center of the data in terms of the column space of the independent variables (see Regression Analysis Problem Set).

We display the data in terms of the independent variables and highlight the high-leverage cases.

```
> # Refit the model using argument x=TRUE so that the lm object includes the
> # matrix of independent variables
> lmfit1<-lm(r.daily.XOM.0 ~ r.daily.SP500.0 + r.daily.DCOILWTICO,
+            data=r.daily.data00,
+            x=TRUE)
> names(lmfit1)

 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "na.action"     "xlevels"       "call"          "terms"
[13] "model"         "x"

> dim(lmfit1$x)

[1] 3371    3

> head(lmfit1$x)

           (Intercept) r.daily.SP500.0 r.daily.DCOILWTICO
2000-01-05           1    0.0017692801       -0.036251729
2000-01-06           1    0.0008049796        0.005663446
2000-01-07           1    0.0265805020        0.000000000
2000-01-10           1    0.0106762566       -0.003232326
2000-01-11           1   -0.0132994562        0.038893791
2000-01-12           1   -0.0045473564        0.023467128
```

We now compute the leverage (and other influence measures) with the function $influence.measures()$ and display the scatter plot of the independent variables, highlighting the high-leverage cases.

```
> lmfit1.inflm<-influence.measures(lmfit1)
> index.inf.hat<-which(lmfit1.inflm$is.inf[,"hat"]==TRUE)
> par(mfcol=c(1,1))
> plot(lmfit1$x[,2], lmfit1$x[,3],xlab="r.daily.SP500.0", ylab="r.daily.DCOILWTICO.0")
> title(main="Scatter Plot of Independent Variables \n High Leverage Points (red o s)")
> points(lmfit1$x[index.inf.hat,2], lmfit1$x[index.inf.hat,3],
+        col=2,
+        pch="o")
>
```


Scatter Plot of Independent Variables
High Leverage Points (red o s)
r.daily.DCOILWTICO.0
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.10
−0.05
0.00
0.05
0.10
r.daily.SP500.0

1.6 References

Lintner, J. (1965). "The Valuation of Risky Assets and the Selection of Risky Investments in Stock Portfolio and Capital Budgets," *Review of Economics and Statistics,* **47:** 13-37.

Sharpe, W. (1964). "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance,* **19:** 425-442.