1. Fit a linear regression model for Ford Motor Co. daily returns using Standard and Poor's 500 Index (SPY) returns as the predictor variable. The slope of that regression (called the *beta* of the company) measures how sensitive the stock's return is to changes in the returns of the overall market (measured by SPY). If the slope is greater than 1, we say that the stock is more volatile than the market. For instance, if the slope is 1.5, then a 1% increase in the SPY index, would result in an average increase of 1.5% in the stock's return.

   The $R^2$ measures the proportion of the total risk that is market-related. For instance, if $R^2 = 0.4$ we would conclude that 40% of the variation in Ford returns is explained by the variation in SPY returns (market-related risk). The remaining 60%, is the proportion of risk that is specific to Ford, and not market-related (firm-specific risk).

   If the market is expected to rise an investor would seek companies with large betas. If market is expected to fall companies with small betas are preferable.

   a) Download daily prices from Ford Motor Co. and Standard and Poor's 500 Index (SPY) from 01-01-2015 to 01-01-2017. Report first 3 and last 3 rows of each.

   b) Find the daily returns for each of them. Use `head()` and `tail()` to report first and last six rows of each.

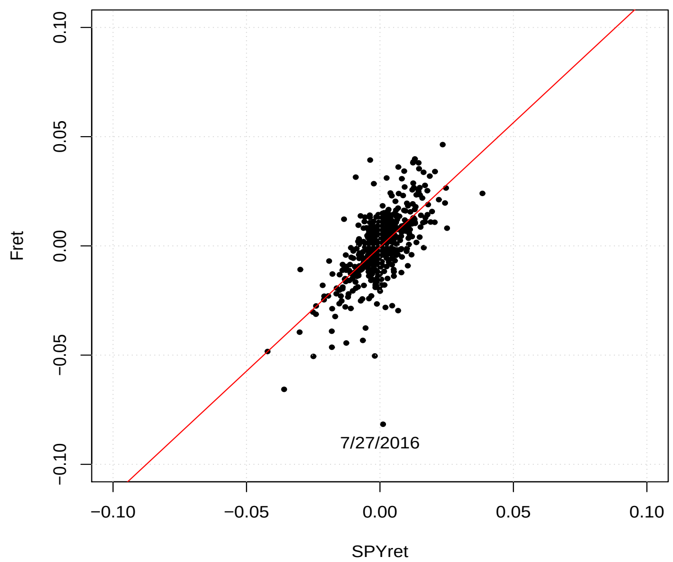
   c) Fit a linear regression model. What is the *beta* of Ford Motor Co.? Interpret Ford's $R^2$.

   d) Create a scatterplot of Ford ($y$-axis) vs. SPY ($x$-axis) daily returns. Show the fitted equation as a dash line on the scatterplot.

   e) Identify the day of the largest outlier. Label that day on the outlier in the scatterplot.

```
# beta.r    F,SPY prices from 01-01-2015 to 01-01-2017

# a) download data
#============================================================================
# Ford
d1= read.csv("F.csv",header=T)
d2= read.csv("SPY.csv",header=T)
n = nrow(d1)
d1[1:3,]
#      Date  Open  High   Low Close Adj.Close   Volume
#1 1/2/2015 15.59 15.65 15.18 15.36  13.22856 24777900
#2 1/5/2015 15.12 15.13 14.69 14.76  12.71181 44079700
#3 1/6/2015 14.88 14.90 14.38 14.62  12.59124 32981600
d1[(n-2):n,]
#          Date  Open  High   Low Close Adj.Close   Volume
#502 12/28/2016 12.37 12.45 12.22 12.25  11.74456 26875400
#503 12/29/2016 12.25 12.31 12.22 12.23  11.72539 19819100
#504 12/30/2016 12.24 12.28 12.08 12.13  11.62952 27405700

# SPY
d2[1:3,]
#        Date   Open   High    Low  Close Adj.Close    Volume
#1 2015-01-02 206.38 206.88 204.18 205.43  195.2395 121465900
#2 2015-01-05 204.17 204.37 201.35 201.72  191.7136 169632600
#3 2015-01-06 202.09 202.72 198.86 199.82  189.9079 209151400
d2[(n-2):n,]
#          Date   Open   High    Low  Close Adj.Close    Volume
#502 2016-12-28 226.57 226.59 224.27 224.40  222.3437  64095000
#503 2016-12-29 224.48 224.89 223.84 224.35  222.2942  48696100
#504 2016-12-30 224.73 224.83 222.73 223.53  221.4817 108998300

# b) Returns
#============================================================================

Fprice = d1$Adj.Close
Fret = Fprice[2:n]/Fprice[1:(n-1)] - 1
head(Fret)
#[1] -0.0390626128 -0.0094849555  0.0287275824  0.0252659624 -0.0136185905  0.0006575888

SPYprice = d2$Adj.Close
SPYret = SPYprice[2:n]/SPYprice[1:(n-1)] - 1
head(SPYret)
# [1] -0.018059635 -0.009418873  0.012461074  0.017745094 -0.008013573 -0.007833607

# c) linear regression
#============================================================================

m1 = lm(Fret~SPYret)
summary(m1)

# Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
```

```
# (Intercept) -0.0004716  0.0005007  -0.942    0.347
# SPYret       1.1374942  0.0556113  20.454   <2e-16 ***

# Residual standard error: 0.01122 on 501 degrees of freedom
# Multiple R-squared:  0.4551,    Adjusted R-squared:  0.454
# F-statistic: 418.4 on 1 and 501 DF,  p-value: < 2.2e-16

# beta is the slope, it is measure of stock's risk
# beta can also be found as follows
cov(SPYret,Fret)/var(SPYret)
# [1] 1.137494

# yahoo reports beta using monthly returns during 3 years

# 45.5% is market-related risk
# 59.5% is risk specific to Ford


# scatterplot
#=========================================================================

plot(Fret~SPYret,pch=19,cex=0.6,xlim=c(-.1,.1),ylim=c(-.1,0.1))
grid()
abline(m1,col="red")

identify(Fret~SPYret,labels=d1$Date)
text(0.0,-0.09,"7/27/2016")
```



2. (30 pts.) Life insurance companies are keenly interested in predicting how long their customers will live because their premiums and profitability depend on such numbers. An actuary gathered data from 100 recently deceased male customers. He recorded the age at death, whether he was a smoker (1 for smoker, 0 for non-smoker), plus the ages at death of his mother and father, the mean ages at death of his grandmothers and grandfathers (see file `insurance.csv`).

   a) Fit a regression model `m1` with all predictors. Use `m2=stepAIC(m1)` to simplify the model. For `m2`

      i. Find a 90% CI on the mean longevity of smokers whose mothers lived to 75 years, whose fathers lived to 65 years, whose grandmothers averaged 85 years, and whose grandfathers averaged 75 years.

      ii. Use `set.seed(2)` to divide the data set into a training and a test set (50%). Compare the $\sqrt{MSPE}$ of `m1` and `m2`.

```
library(MASS)     # stepAIC()

d0=read.csv("insurance.csv")
str(d0)
# 'data.frame':   100 obs. of  6 variables:
#  $ Longevity: int  80 73 70 72 79 83 70 72 72 71 ...
#  $ Mother   : int  85 88 66 72 88 90 67 76 66 78 ...
#  $ Father   : int  78 63 75 67 73 72 65 71 75 64 ...
#  $ Gmothers : int  72 76 67 68 64 74 70 74 71 76 ...
#  $ Gfathers : int  71 66 57 55 73 62 59 61 63 61 ...
#  $ Smoker   : int  0 1 1 1 0 0 1 1 1 1 ...

# Smoker is a factor
d0$Smoker = factor(d0$Smoker)

# a) full model
m1 = lm(Longevity~.,d0)
summary(m1)
# Coefficients:
#            Estimate Std. Error t value Pr(>|t|)
#(Intercept) 23.56735    5.97848   3.942 0.000155 ***
#Mother       0.30612    0.05420   5.648 1.72e-07 ***
#Father       0.30301    0.04758   6.368 6.99e-09 ***
#Gmothers     0.03161    0.05772   0.548 0.585286
#Gfathers     0.07779    0.05729   1.358 0.177712
#Smoker1     -3.71899    0.66908  -5.558 2.54e-07 ***

#Residual standard error: 2.323 on 94 degrees of freedom
#Multiple R-squared:  0.8051,    Adjusted R-squared:  0.7947
#F-statistic: 77.66 on 5 and 94 DF,  p-value: < 2.2e-16
```

```
# simplified model
m2 = stepAIC(m1)
summary(m2)

# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) 27.22780    5.09604   5.343 6.15e-07 ***
# Mother       0.33444    0.04747   7.046 2.79e-10 ***
# Father       0.32376    0.04483   7.222 1.21e-10 ***
# Smoker1     -3.73771    0.66732  -5.601 2.03e-07 ***

# Residual standard error: 2.322 on 96 degrees of freedom
# Multiple R-squared:  0.8012,    Adjusted R-squared:  0.795
# F-statistic:   129 on 3 and 96 DF,  p-value: < 2.2e-16

# Longevity of parents and smoker status explain 80.12% of customer longevity
# For each additional year Mother lived longevity increases 0.33, c.p.
# For each additional year Father lived longevity increases 0.32, c.p.
# Smokers live 3.73 less years than non-smokers, on average

# b) prediction
#=====================================================================

newval = subset(d0,select=c(Mother,Father,Smoker))[1,]
newval$Mother = 75
newval$Father = 65
newval[1,3]=1
str(newval)
predict(m2,newval,interval="conf")
#       fit      lwr      upr
# 1 69.61732 68.99821 70.23644
```

```
# c) validation
#=======================================================================

set.seed(2)
train = sample(100,50)

# mspe for reduced model

d2 = d0[,-c(4,5)]
dtrain = d2[train,]
dtest  = d2[-train,]
ytest = dtest$Longevity         # response variable from dtest set

m2b = lm(Longevity~.,dtrain)
yhat2 = predict(m2b,dtest)       # predictions on dtest set

a = mean((ytest-yhat2)^2)      # 6.391106
sqrt(a)                        # 2.519974 years away from perfect prediction, on average

# mspe for full model

m1b = lm(Longevity~.,d0[train,])
yhat1 = predict(m1b,d0[-train,])       # predictions on dtest set
# mspe
b = mean((ytest-yhat1)^2)      # 6.391106
sqrt(b)                        # 2.587677 years away from perfect prediction, on average

# model with less predictors with better prediction performance than
# model with all predictors
```

3. (30 pts.) The data set `stockdata` from library `huge` consists of the price and company information of S&P 500 stock shares. The data set consists of two dataframes `names.csv` and `prices.csv`. They are available on Blackboard.

   a) Find the correlation matrix `C` of these prices. Then use `which(C==max(C),arr.ind=T)` to find the largest correlation, and their row and column numbers in `C`. Identify the companies with the largest correlation. Report their full name.

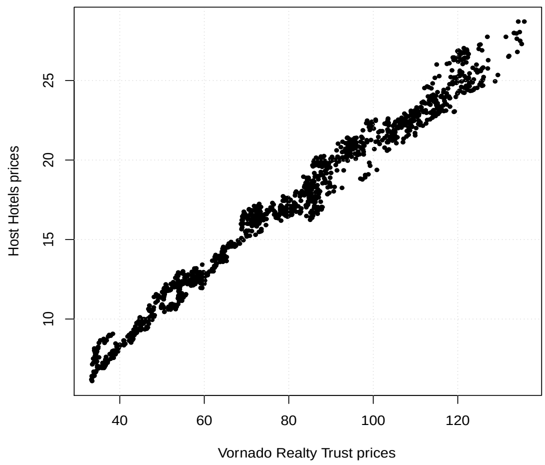
   b) Build a scatterplot of their prices

   c) How many `Health Care` companies are in the full dataset?

   d) Report the name of the two most correlated companies in the Financial Sector.

```
d0=read.csv("prices.csv",header=T)
names0=read.csv("names.csv",header=T)

# head(names0)
#  Ticker                 Sector                          Name
#1    MMM            Industrials                         3M Co
#2    ACE             Financials                   ACE Limited
#3    ABT            Health Care          Abbott Laboratories
#4    ANF Consumer Discretionary Abercrombie & Fitch Company A
#5   ADBE Information Technology            Adobe Systems Inc
#6    AMD Information Technology       Advanced Micro Devices
str(names0)
#'data.frame':   452 obs. of  3 variables:
# $ Ticker: Factor w/ 452 levels "A","AA","AAPL",..: 271 6 5 30 7 25 14 15 16 1 ...
# $ Sector: Factor w/ 10 levels "Consumer Discretionary",..: 6 4 5 1 7 7 10 5 4 7 ...
# $ Name  : Factor w/ 452 levels "3M Co","Abbott Laboratories",..: 1 4 2 3 5 6 7 8 9 10 .


# correlation
#========================================================================
a=cor(d0)
diag(a)=0
which(a == max(a), arr.ind = T)

#    row col
#VNO 428 205
#HST 205 428
a[205,428]
# [1] 0.9901256

# Companies with largest correlation

names0[428,]
#     Ticker     Sector                Name
# 428    VNO Financials Vornado Realty Trust
names0[205,]
#     Ticker     Sector                 Name
# 205    HST Financials Host Hotels & Resorts
```

```
# plot their prices
#==========================================================================

VNO = d0[,428]
HST = d0[,205]
plot(HST~VNO)
plot(HST~VNO,pch=19,cex=0.6,xlab="Vornado Realty Trust prices",ylab="Host Hotels prices")
grid()

# companies per sector
#==========================================================================

table(names0$Sector)

#     Consumer Discretionary            Consumer Staples
#                         70                          35
#                     Energy                  Financials
#                         37                          74
#                Health Care                 Industrials
#                         46                          59
#     Information Technology                   Materials
#                         64                          29
#Telecommunications Services                   Utilities
#                          6                          32

# Financial companies
#==========================================================================

names1 = names0[names0$Sector == "Financials",]
dim(names1)
# [1] 74  3
head(names1)
#   Ticker     Sector                   Name
#2     ACE Financials            ACE Limited
#9     AFL Financials              AFLAC Inc
#18    ALL Financials          Allstate Corp
#24    AXP Financials     American Express Co
#25    AIG Financials American Intl Group Inc
#32    AON Financials        Aon Corporation
tail(names1)
#     Ticker     Sector                 Name
#420    UNM Financials           Unum Group
#425    VTR Financials           Ventas Inc
#428    VNO Financials Vornado Realty Trust
#437    WFC Financials          Wells Fargo
#448     XL Financials           XL Capital
#452   ZION Financials        Zions Bancorp
str(names1)
# 'data.frame':   74 obs. of  3 variables:
#  $ Ticker: Factor w/ 452 levels "A","AA","AAPL",..: 6 16 22 42 18 31 19 39 45 56 ...
#  $ Sector: Factor w/ 10 levels "Consumer Discretionary",..: 4 4 4 4 4 4 4 4 4 4 ...
#  $ Name  : Factor w/ 452 levels "3M Co","Abbott Laboratories",..: 4 9 18 24 25 32 34 44
```




```
idx = names1$Ticker
idx = as.character(idx)
head(idx)
# [1] "ACE" "AFL" "ALL" "AXP" "AIG" "AON"
tail(idx)
# [1] "UNM"  "VTR"  "VNO"  "WFC"  "XL"   "ZION"

# prices of Financials
d1=d0[,idx]
dim(d1)
# [1] 1258   74

# correlation of Financials
B=cor(d1)
diag(B)=0
which(B == max(B), arr.ind = T)

#     row col
# VNO  71  32
# HST  32  71

B[32,71]
# [1] 0.9901256

# same companies as in part (a)

names0[names0$Ticker=="VNO",]
#     Ticker     Sector                 Name
# 428    VNO Financials Vornado Realty Trust
names0[names0$Ticker=="HST",]
#     Ticker     Sector                  Name
# 205    HST Financials Host Hotels & Resorts
```