46-926 Homework 2, Part II

Jingyi Guo, Pittsburgh Campus

1/28/2017

Preparation 1

```
fullrow=rep(FALSE,nrow(trainset))
for (i in 1:nrow(trainset))
{
  fullrow[i]=!any(is.na(trainset[i,29:147]))
}
```

Now 40000 logical values(i.e. TRUE or FALSE) are stored in fullrow.

Preparation 2

```
varnames <- c(paste("Ret_", 2:120, sep=""))
fullform = as.formula(paste("Ret_PlusOne ~ ",paste(varnames,collapse="+")))
print(fullform)
```

```
## Ret_PlusOne ~ Ret_2 + Ret_3 + Ret_4 + Ret_5 + Ret_6 + Ret_7 +
##     Ret_8 + Ret_9 + Ret_10 + Ret_11 + Ret_12 + Ret_13 + Ret_14 +
##     Ret_15 + Ret_16 + Ret_17 + Ret_18 + Ret_19 + Ret_20 + Ret_21 +
##     Ret_22 + Ret_23 + Ret_24 + Ret_25 + Ret_26 + Ret_27 + Ret_28 +
##     Ret_29 + Ret_30 + Ret_31 + Ret_32 + Ret_33 + Ret_34 + Ret_35 +
##     Ret_36 + Ret_37 + Ret_38 + Ret_39 + Ret_40 + Ret_41 + Ret_42 +
##     Ret_43 + Ret_44 + Ret_45 + Ret_46 + Ret_47 + Ret_48 + Ret_49 +
##     Ret_50 + Ret_51 + Ret_52 + Ret_53 + Ret_54 + Ret_55 + Ret_56 +
##     Ret_57 + Ret_58 + Ret_59 + Ret_60 + Ret_61 + Ret_62 + Ret_63 +
##     Ret_64 + Ret_65 + Ret_66 + Ret_67 + Ret_68 + Ret_69 + Ret_70 +
##     Ret_71 + Ret_72 + Ret_73 + Ret_74 + Ret_75 + Ret_76 + Ret_77 +
##     Ret_78 + Ret_79 + Ret_80 + Ret_81 + Ret_82 + Ret_83 + Ret_84 +
##     Ret_85 + Ret_86 + Ret_87 + Ret_88 + Ret_89 + Ret_90 + Ret_91 +
##     Ret_92 + Ret_93 + Ret_94 + Ret_95 + Ret_96 + Ret_97 + Ret_98 +
##     Ret_99 + Ret_100 + Ret_101 + Ret_102 + Ret_103 + Ret_104 +
##     Ret_105 + Ret_106 + Ret_107 + Ret_108 + Ret_109 + Ret_110 +
##     Ret_111 + Ret_112 + Ret_113 + Ret_114 + Ret_115 + Ret_116 +
##     Ret_117 + Ret_118 + Ret_119 + Ret_120
```

Now, fullform is a formula that can be used for regression.

Fit Linear Model

```
fitmodel=lm(fullform,data=subset(trainset,fullrow==TRUE))
summary(fitmodel)
```

```
##
## Call:
## lm(formula = fullform, data = subset(trainset, fullrow == TRUE))
```

```
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.25214 -0.01128 -0.00010  0.01104  0.40607 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.0001893  0.0001754  -1.079 0.280491    
## Ret_2       -0.1629095  0.1710270  -0.953 0.340835    
## Ret_3       -0.0607402  0.1662972  -0.365 0.714928    
## Ret_4       -0.9347786  0.1688496  -5.536 3.13e-08 ***
## Ret_5        0.2654575  0.1707327   1.555 0.120005    
## Ret_6       -0.4152042  0.1792979  -2.316 0.020582 *  
## Ret_7       -0.3043872  0.1703990  -1.786 0.074061 .  
## Ret_8       -0.8974015  0.1767239  -5.078 3.85e-07 ***
## Ret_9       -0.7741098  0.1766342  -4.383 1.18e-05 ***
## Ret_10      -0.9008347  0.1684927  -5.346 9.06e-08 ***
## Ret_11      -0.1620638  0.1683873  -0.962 0.335836    
## Ret_12       0.6958409  0.1721282   4.043 5.30e-05 ***
## Ret_13      -0.3585545  0.1800424  -1.992 0.046438 *  
## Ret_14      -0.5718430  0.1724288  -3.316 0.000913 ***
## Ret_15       0.4279127  0.1717690   2.491 0.012738 *  
## Ret_16      -0.4756241  0.1743516  -2.728 0.006378 ** 
## Ret_17      -0.4025796  0.1744930  -2.307 0.021056 *  
## Ret_18      -0.9117559  0.1718611  -5.305 1.14e-07 ***
## Ret_19       0.4166300  0.1766885   2.358 0.018383 *  
## Ret_20      -0.6280223  0.1797067  -3.495 0.000476 ***
## Ret_21      -0.2568366  0.1773856  -1.448 0.147659    
## Ret_22      -0.8902819  0.1715447  -5.190 2.12e-07 ***
## Ret_23       0.5885632  0.1782561   3.302 0.000962 ***
## Ret_24       0.3664450  0.1735488   2.111 0.034742 *  
## Ret_25       0.5485325  0.1797646   3.051 0.002280 ** 
## Ret_26       0.6688867  0.1755138   3.811 0.000139 ***
## Ret_27       1.3162605  0.1763514   7.464 8.71e-14 ***
## Ret_28      -0.6329628  0.1738220  -3.641 0.000272 ***
## Ret_29       1.9163135  0.1730064  11.077  < 2e-16 ***
## Ret_30      -0.0022790  0.1832110  -0.012 0.990075    
## Ret_31      -0.7435309  0.1812794  -4.102 4.12e-05 ***
## Ret_32      -0.0887892  0.1667780  -0.532 0.594469    
## Ret_33      -0.4120041  0.1735593  -2.374 0.017612 *  
## Ret_34      -1.1473586  0.1670472  -6.868 6.66e-12 ***
## Ret_35       0.1221068  0.1726814   0.707 0.479498    
## Ret_36      -0.8441781  0.1771196  -4.766 1.89e-06 ***
## Ret_37      -1.1055341  0.1712583  -6.455 1.10e-10 ***
## Ret_38      -0.2737043  0.1742205  -1.571 0.116192    
## Ret_39      -0.0409682  0.1752295  -0.234 0.815144    
## Ret_40       0.2760586  0.1730336   1.595 0.110636    
## Ret_41       0.2976071  0.1665179   1.787 0.073913 .  
## Ret_42      -0.4057437  0.1745787  -2.324 0.020127 *  
## Ret_43       0.5478472  0.1739782   3.149 0.001641 ** 
## Ret_44       0.8566940  0.1672313   5.123 3.04e-07 ***
## Ret_45       0.2357886  0.1781661   1.323 0.185709    
## Ret_46      -0.1460974  0.1734107  -0.842 0.399521    
## Ret_47      -0.1421768  0.1740072  -0.817 0.413895    
## Ret_48       0.4736266  0.1703059   2.781 0.005423 **
## Ret_49      -0.4687539  0.1750756  -2.677 0.007424 **
## Ret_50      -1.1867554  0.1802126  -6.585 4.64e-11 ***
## Ret_51       0.0103413  0.1679979   0.062 0.950917
## Ret_52       0.2047506  0.1669989   1.226 0.220189
## Ret_53       0.7911645  0.1712637   4.620 3.87e-06 ***
## Ret_54       0.0963189  0.1727107   0.558 0.577062
## Ret_55      -0.7851174  0.1765537  -4.447 8.75e-06 ***
## Ret_56      -0.1096328  0.1579839  -0.694 0.487721
## Ret_57       1.1264546  0.1675770   6.722 1.84e-11 ***
## Ret_58       0.2504568  0.1586010   1.579 0.114313
## Ret_59      -0.4423694  0.1783457  -2.480 0.013131 *
## Ret_60       0.7018259  0.1805171   3.888 0.000101 ***
## Ret_61       0.6354898  0.1692577   3.755 0.000174 ***
## Ret_62      -1.2025635  0.1571365  -7.653 2.04e-14 ***
## Ret_63      -0.0824153  0.1607284  -0.513 0.608124
## Ret_64      -0.8723008  0.1662265  -5.248 1.55e-07 ***
## Ret_65      -0.4696331  0.1661103  -2.827 0.004699 **
## Ret_66       0.9116911  0.1692801   5.386 7.29e-08 ***
## Ret_67      -0.0164970  0.1553383  -0.106 0.915424
## Ret_68       0.3444063  0.1485333   2.319 0.020420 *
## Ret_69      -0.3583904  0.1704715  -2.102 0.035534 *
## Ret_70       0.3747183  0.1676870   2.235 0.025452 *
## Ret_71      -0.8002245  0.1658033  -4.826 1.40e-06 ***
## Ret_72      -1.0649038  0.1723210  -6.180 6.53e-10 ***
## Ret_73      -0.5573758  0.1656255  -3.365 0.000766 ***
## Ret_74       0.8902702  0.1768475   5.034 4.84e-07 ***
## Ret_75       1.0599664  0.1721721   6.156 7.57e-10 ***
## Ret_76       0.0764507  0.1359614   0.562 0.573919
## Ret_77       0.2723177  0.1449624   1.879 0.060320 .
## Ret_78       0.4079790  0.1404106   2.906 0.003669 **
## Ret_79       0.4303133  0.1582256   2.720 0.006541 **
## Ret_80      -0.7425012  0.1555995  -4.772 1.84e-06 ***
## Ret_81      -0.4292205  0.1582941  -2.712 0.006702 **
## Ret_82      -0.2870886  0.1549810  -1.852 0.063980 .
## Ret_83       0.3329457  0.1627507   2.046 0.040794 *
## Ret_84       0.6393141  0.1589273   4.023 5.77e-05 ***
## Ret_85      -0.9824220  0.1561736  -6.291 3.22e-10 ***
## Ret_86      -1.1304205  0.1541267  -7.334 2.30e-13 ***
## Ret_87      -0.3379539  0.1415025  -2.388 0.016934 *
## Ret_88       0.1983830  0.1625947   1.220 0.222437
## Ret_89      -0.4069849  0.1601028  -2.542 0.011028 *
## Ret_90       0.7129373  0.1546626   4.610 4.06e-06 ***
## Ret_91      -0.3520522  0.1575628  -2.234 0.025469 *
## Ret_92       0.3246340  0.1511557   2.148 0.031750 *
## Ret_93       0.1399515  0.1556647   0.899 0.368632
## Ret_94      -0.0804870  0.1462473  -0.550 0.582086
## Ret_95      -0.2246888  0.1569362  -1.432 0.152238
## Ret_96       0.1227909  0.1586220   0.774 0.438874
## Ret_97      -0.0911997  0.1488519  -0.613 0.540089
## Ret_98      -0.1193954  0.1496839  -0.798 0.425082
## Ret_99      -0.5988275  0.1570194  -3.814 0.000137 ***
## Ret_100     -0.8463234  0.1367994  -6.187 6.25e-10 ***
## Ret_101     -1.7576910  0.1495219 -11.755  < 2e-16 ***
## Ret_102     -0.9254008  0.1565992  -5.909 3.48e-09 ***
## Ret_103     -0.9054849  0.1650569  -5.486 4.16e-08 ***
## Ret_104      0.5005022  0.1479920   3.382 0.000721 ***
## Ret_105     -1.3500325  0.1686634  -8.004 1.26e-15 ***
## Ret_106     -0.1668115  0.1655209  -1.008 0.313563
## Ret_107     -1.5557886  0.1605753  -9.689  < 2e-16 ***
## Ret_108      0.1279454  0.1481396   0.864 0.387772
## Ret_109     -0.7343552  0.1535756  -4.782 1.75e-06 ***
## Ret_110      0.0926245  0.1538253   0.602 0.547086
## Ret_111      0.1732656  0.1615340   1.073 0.283450
## Ret_112      0.2230703  0.1561864   1.428 0.153239
## Ret_113     -0.8822088  0.1499731  -5.882 4.10e-09 ***
## Ret_114     -0.3159368  0.1624329  -1.945 0.051784 .
## Ret_115     -1.1343207  0.1664150  -6.816 9.59e-12 ***
## Ret_116     -0.6790886  0.1584757  -4.285 1.83e-05 ***
## Ret_117     -0.5611371  0.1525447  -3.679 0.000235 ***
## Ret_118     -0.0760989  0.1611384  -0.472 0.636747
## Ret_119     -0.7931603  0.1551260  -5.113 3.20e-07 ***
## Ret_120      0.4039446  0.1492010   2.707 0.006787 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02537 on 22270 degrees of freedom
## Multiple R-squared:  0.1547, Adjusted R-squared:  0.1501
## F-statistic: 34.24 on 119 and 22270 DF,  p-value: < 2.2e-16
```

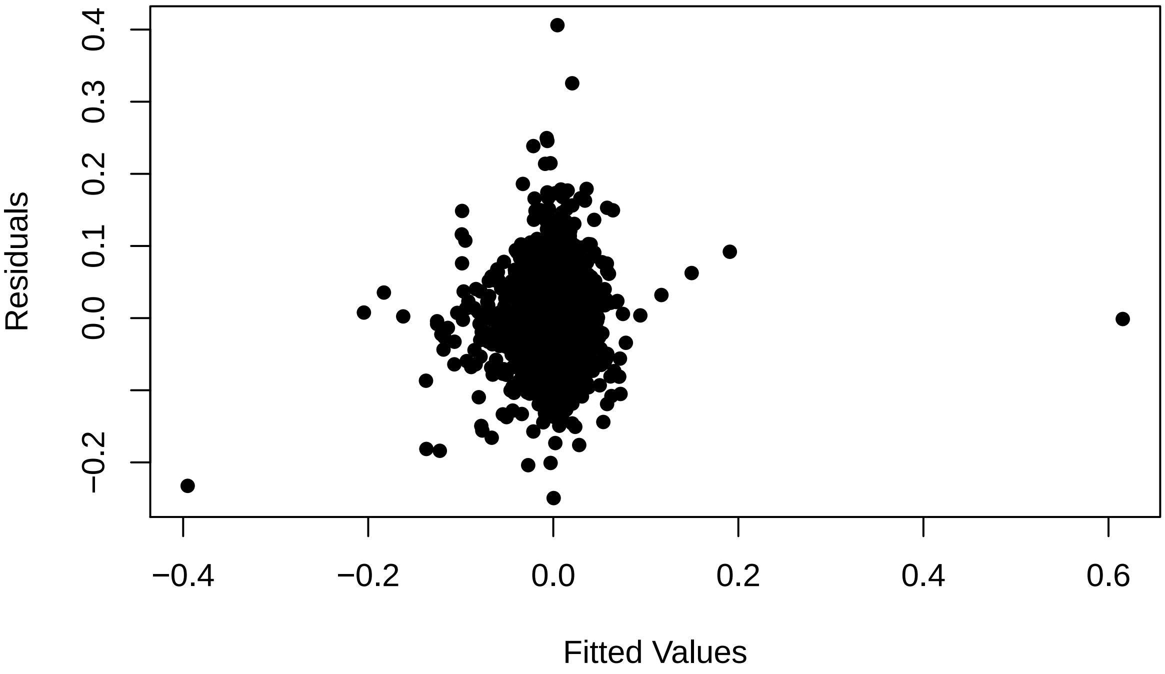
Stepwise Variable Selection

```
finalmod=step(fitmodel,direction="both")
```

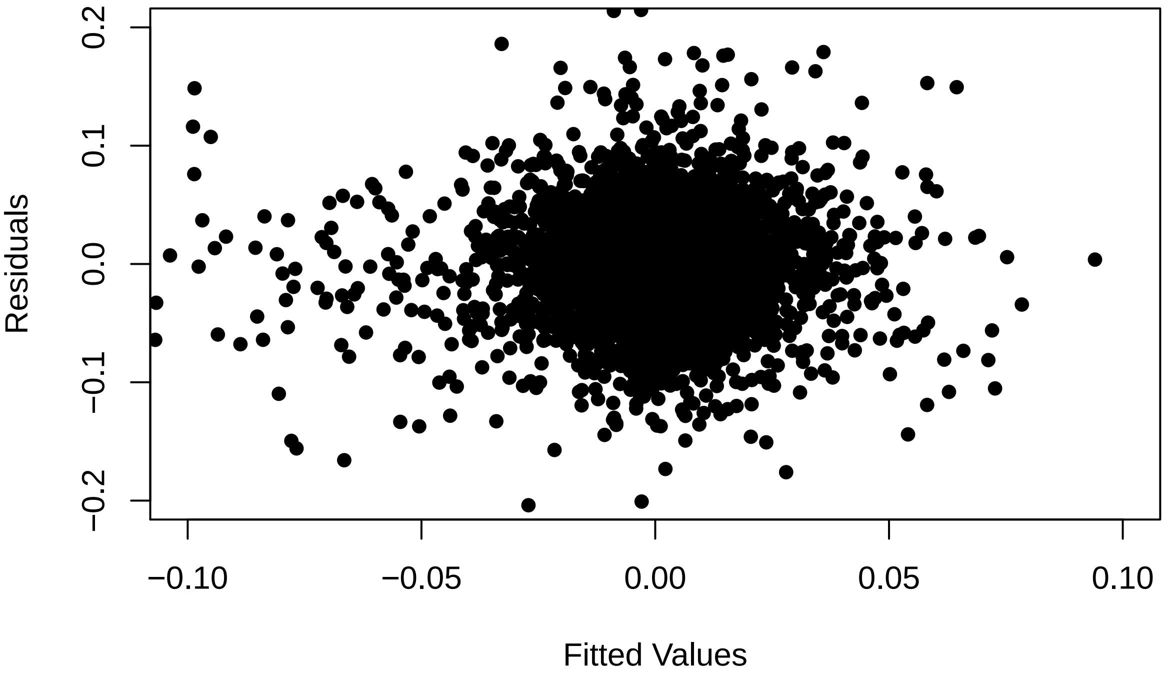
(The result is hidden due to length) 90 predictors are retained: Ret_4,Ret_5,Ret_6,Ret_7, Ret_8,Ret_9,Ret_10,Ret_12, Ret_13,Ret_14,Ret_15,Ret_16, Ret_17,Ret_18,Ret_19,Ret_20, Ret_21,Ret_22,Ret_23,Ret_24, Ret_25,Ret_26,Ret_27,Ret_28, Ret_29,Ret_31,Ret_33,Ret_34, Ret_36,Ret_37,Ret_38,Ret_40, Ret_41,Ret_42,Ret_43,Ret_44, Ret_48,Ret_49,Ret_50,Ret_53, Ret_55,Ret_57,Ret_58,Ret_59, Ret_60,Ret_61,Ret_62,Ret_64, Ret_65,Ret_66,Ret_68,Ret_69, Ret_70,Ret_71,Ret_72,Ret_73, Ret_74,Ret_75,Ret_77,Ret_78, Ret_79,Ret_80,Ret_81,Ret_82, Ret_83,Ret_84,Ret_85,Ret_86, Ret_87,Ret_89,Ret_90,Ret_91, Ret_92,Ret_95,Ret_99,Ret_100, Ret_101,Ret_102,Ret_103,Ret_104, Ret_105,Ret_107,Ret_109,Ret_113, Ret_114,Ret_115,Ret_116,Ret_117, Ret_119,Ret_120

Residual plot

```
plot(as.numeric(finalmod$fit),as.numeric(finalmod$resid),pch=16,xlab="Fitted Values", ylab="Residuals")
```



```
plot(as.numeric(finalmod$fit),as.numeric(finalmod$resid),pch=16,xlab="Fitted Values", ylab="Residuals",
```



Comment: There is no significantly prevalent pattern in the plot of residual versus fitted values. However, there are some points for which the residual is quite extreme relative to others.

Cook's Distance

```
cookd=as.numeric(cooks.distance(finalmod))
sort(pf(cookd,91,22299),decreasing=TRUE)[1:5]
```

```
## [1] 1.000000e+00 9.999998e-01 4.866206e-06 3.670158e-14 7.655050e-21
```

The largest two Cook's Distance's exceed the median of the F distribution, so they are definitely cause for concern as being too influential.