Determinants of Mutual Fund Performance

Nathan Rule, Miles Carpenter, and Thomas Murawski

Executive Summary

Mutual funds are a very important and distinct segment of the financial market. They vary greatly in size, investment strategies, and general structure. Although it is impossible to perfectly predict future outcomes in the financial world, any indication of future returns could be very helpful to investors and managers alike. This report uses a collection of mutual funds to create a regression model that explains funds year to date returns in terms of annual holdings turnover, worst 3-year return, and fund valuation. In addition to other information, the model indicated that funds with higher annual holdings turnover and lower worst 3-year returns had higher year to date returns. Also, growth funds were expected to have higher returns than either blend or value funds. This model, while giving some insight into which determinants affect fund performance, could also be used to predict future performance of similar mutual funds.

Section 1. Introduction

Mutual funds are professionally managed collections of stocks, bonds and other securities. Money is pooled from many and invested by a fund manager. The fund manager trades the fund's underlying securities, realizes capital gains or losses, and collects the dividend or interest income from the assets. The investment proceeds are then passed along to the individual investors. In exchange for managing and maintaining the mutual fund, the manager charges a fee which is deducted from the shareholders' earnings. Money is invested in a mutual fund by purchasing shares of the fund. Mutual fund shares are analogous to shares of stock, as the shareholders are considered to be owners of the fund. Shareholders have voting rights in proportion to their ownership of the fund.

The first mutual fund was the Massachusetts Investors Trust founded on March 21, 1924. After one year, the fund had 200 shareholders and $392,000 in assets. The mutual fund industry is growing extremely rapidly. There are now more mutual funds than stocks on the New York Stock Exchange. A major contributor of mutual fund growth was the provision added to the Internal Revenue Code in 1975 that allows individuals to contribute $2,000 a year to their individual retirement accounts (IRA). Mutual funds are now popular in employer-sponsored retirement plans, IRAs and Roth IRAs.

As of October 2007, there are 8,015 mutual funds that belong to the Investment Company Institute (ICI) with combined assets of $12.356 trillion. The ICI is an American investment trade organization. According to its website, the ICI is responsible for "encouraging adherence to high ethical standards by all industry participants; advancing the interests of funds, their shareholders, directors, and investment advisers; and promoting public understanding of mutual funds and other investment companies."

The nature of mutual funds allows them to invest in different kinds of securities. The most common are cash, stock, and bonds. For the purpose of this project, the mutual funds that were analyzed held varying amounts of stock and cash.

Investors often conduct light research in order to determine which mutual fund is the best to invest in. There are many sources available that report mutual fund performance over time. The concept of portfolio performance has two dimensions: the ability of the portfolio to minimize risk through efficient diversification, and the "ability of the portfolio to increase returns through successful prediction of future security prices."(Jensen) Predicting future prices is extremely difficult. On average, mutual funds have little to no ability to forecast the market. Approximately 80% of all mutual funds under perform the average return of the stock market after management fees are deducted.

Mutual Fund returns are affected by numerous factors. The types of assets a fund owns will impact its earnings. More specifically, a fund's objective can affect results. For example, a fund can invest in a specific industry, such as technology. Oftentimes, a fund will also concentrate on investing in growth or income stocks. Other metrics, such as asset turnover, expense ratio, and standard deviation may have an impact on earnings.

The data for this study was found on the Yahoo! Finance website. Yahoo! maintains an extensive database of daily returns and other facts for many securities. The goal of this report is to study the year to date returns of randomly selected mutual funds and investigate which factors have a significant impact. Based on this, we will construct various models and test their assumptions to determine their worth.

In the remainder of the paper, variables will be chosen, models developed and tested, and final conclusions reached. Section 2 outlines the data we gathered. Section 3 provides information about the models that were developed, with most of the rigorous technical work in the appendices. The results of the report, as well as possible limitations and improvements, can be found in Section 4.

Section 2. Data characteristics

In order to get a well-rounded sample to create a model from, mutual funds were selected randomly from random families of funds using the online tools of Yahoo finance. After the funds were chosen, variables were then collected that seemed to sum up the performance and composition of the funds. The response variable, annualized year to date return was collected for the funds, along with other explanatory variables. The variables collected are summarized in the appendix.

The data collected represents funds in all sectors of the market, with widely varied returns, investments, strategies and sizes. The response variable, year to date return, ranges from a low of -1.61% to a high of 45.4%. The average year to date return is 13.41%. This is more than twice the year to date return of the S&P 500 index, which is 6.28%, so the majority of the funds selected are beating the market for this calendar year. This is useful in terms of the model, because the majority of funds investors would be interested in are usually going to be beating the market regularly. There is much more purpose to predicting the returns of successful funds than unsuccessful ones.

Two qualitative characteristics of the funds that were chosen are size and valuation. These variables correspond to the types of companies the fund is investing in, and the overall aim of the fund. Grouped together, these two categories make up the Morning Star Style Box, which is a useful tool in classifying mutual funds. It was hypothesized that funds valuated as growth

would potentially have higher returns, as their aim is to grow their investments. Also, funds classified as small may be investing in IPO's and small startup companies, which may provide a larger opportunity for higher returns. The following tables show the number of funds which fall into each of these categories, and the average year to date returns for each category:

Average Year-to-Date Returns by Size

Fund Size	Number of Funds	Average YTD Rtn
Small	8	15.43
Medium	6	13.52
Large	24	12.71
Total:	38	13.41

Average Year-to-Date Returns by Valuation

Fund Valuation	Number of Funds	Average YTD Rtn
Value	10	8.29
Blend	11	10.66
Growth	17	18.20
Total:	38	13.41

As predicted, the small funds seem to have a slightly higher year to date return, while medium sized funds mirror the average returns fairly closely. Looking at the fund valuation data, there are some interesting trends. The value funds are returning at much lower rates than the average of the collected funds, almost as low as the market index. The growth funds however, are returning much higher than average, which was predicted. This could be a function of the sample size, or could be a result of fundamental investing strategies. There's no way to know whether these differences in means are significant or not until a model is created and diagnostics are run, but it will be kept in mind during variable selection.

Originally, when the data was collected, a number of funds were hand picked from among the best performing and worst performing funds currently in the market. This was done with the belief that a model based on these observations as well as more 'average' ones would have more universal predictive value. Some of these funds returned as much as 110% this year, and some as low as -60%. This obviously provided a much larger range of data, but forced these hand picked values to almost certainly be outliers. After an initial analysis of the data, it was determined that these values provided too much variance in the observations. Not only was year to date return greatly affected, but standard deviation, annual holdings turnover, best 3-year return, and others were altered as well. It was determined that whatever added predictive value there was became overshadowed by the negative consequences. After careful thought, it was decided that these hand picked values would be excluded in favor of a thoroughly random sample. The rationale was that this would truly represent a random slice of the market, and would be more accurate in predicting the year to date returns of 'average' funds. Despite this exclusion, there are still observations of negative returns and very, very high returns, so not too much variation has been lost.

Section 3. Model selection and interpretation

In rigorously working with the data and testing numerous models we found a decent amount of correlation between characteristics of a mutual fund and the year to date returns of that fund. This section looks at a couple regression models that describes this pattern. The model and its interpretation have been provided here, with motivations and more in-depth analysis in the appendices.

The model we found to yield the best results is as follows:

(1) Predicted YTD rtn = 7.50588 + .07146*AHT + -.2068*worst 3 year rtn + 1.37619*valuationGrowth + -3.53535*valuationValue

Due to the categorical term in this model it actually decomposes into three separate models depending on the valuation of the mutual fund. If the mutual fund is a Blend mutual fund, then the valuationGrowth and valuationValue terms are zero and the model looks like the following:

(2) Predicted YTD rtn = 7.50588 + .07146*AHT + -.2068*worst 3 year rtn

If the mutual fund of the type Value then the valuationValue term is 1 and the valuationGrowth term is zero. The model then looks like:

(3) Predicted YTD rtn = 7.50588 + .07146*AHT + -.2068*worst 3 year rtn + -3.53535*1

If the mutual fund is of the type Growth then the valuationValue term is 0 and the valuationGrowth term is 1. The model then looks as such:

(4) Predicted YTD rtn = 7.50588 + .07146*AHT + -.2068*worst 3 year rtn + 1.37619*1

The dependent variable in the model is the year to date return of the mutual fund. The explanatory variables are the annual holdings turnover (AHT), the worst three-year return over the life of the mutual fund, and the valuation of the mutual fund. The valuation is broken down further into Growth, Value and, imbedded in the intercept, Blend. This model is used to predict the year to date return of any stock based mutual fund. Let's do an example to further explain. American Trust Allegiance mutual fund (ATAFX) has an annual holdings turnover of 80%, a worst 3-yr return of -19.19%, and is a growth mutual fund. Based on this information the model looks as follows:

Predicted YTD rtn = 7.50588 + .07146*80 + (-.2068)*(-19.19) + 1.37619*1 = 18.567362

This has an estimated year to date return of 18.567%. The actual year to date return is 15.23%. Obviously there is a difference in the values; however the model came fairly close just using three values from the mutual fund's financial statements.

The coefficients in model (1) tell a lot about the relationship between each of the different factors and the year to date return. The intercept in the model accounts for the valuationBlend term, so its interpretation is different from a linear model without a categorical value. It now suggests that a blend type mutual fund with zero annual holdings turnover and a 3-yr worst return of zero has an expected year to date return of 7.50588%. This seems to make sense, as it

is slightly higher then the return on government bonds and it is expected that a mutual fund with all its expertise could choose a set of holdings at the beginning of the year that would do better then the risk free rate (government bonds).

The annual holdings turnover coefficient of .07146 suggests that with all else being held constant if a mutual fund increases its annual holdings turnover by 1 percent it would expect its expected year to date return to increase by .07146%.

The worst 3-year return coefficient suggests that if a mutual funds worst 3 year return were 1 percent higher it would expect its year to date return to decrease by .2068 percent, all else held constant. This doesn't seem to make sense at first glance, but if you think about it for a second it does hold water. A lower worst 3-year return suggests a higher standard deviation on the mutual fund, and a high standard deviation leads to the potential for greater returns in this year.

The coefficient for valuationValue is the expected difference in the year to date return for a Blend mutual fund over a Value mutual fund. A coefficient of -3.53535 suggests that a Value mutual fund expects to return about 3.5% lower then a Blend mutual fund. This is intuitive as a value mutual fund has less risk it expects on average to have a lower return.

The coefficient for valuationGrowth is very similar except it is the expected difference in the expected year to date return for a Blend mutual fund verses a Growth mutual fund. A coefficient of 1.37619 makes sense, as a Growth mutual fund should expect on average to earn a higher return then a blend mutual fund, as it invests in riskier assets.

We found the model to fit the data relatively well. Both the intercept, containing valuationValue term, and the annual holdings turnover variables were significant to an alpha .001 or smaller. In other words they were extremely significant. Also the worst 3-yr return variable was significant to an alpha level of 5%, which is very good as well. The model also had an adjusted R-squared value of .6424. This is considerably strong, as 64.24% of variation was explained by the model. In doing diagnostics on the model we found the data to have a few outliers and leverage points. After removing all of these points, the model still had an adjusted R-squared value of .4239. Further diagnostics of the model can be found in the appendix. It outlines where we started with the model and goes stepwise to where we ended. It checks that the residuals were normally distributed, independently identically distributed and had constant variance. It also checks for any collinearity between the variables. In general we were very satisfied with all of the results we found.

After completing what we determined to be the best possible model, we went back to some of the other variables that were more intuitive and created an alternative model. The model looks as follows.

Predicted YTD Returns = 1.58998 + 1.01659*SD + -.31745*worst 3-year rtn

As you can see the only new value in the model is that of standard deviation (SD). It having a positive correlation makes sense intuitively as we would expect a mutual fund with a higher standard deviation or risk to have a higher expected payout or return.

From the summary (appendix) we can see that both of the explanatory variables are significant to an alpha of .01, which is very strong. Also from the adjusted R-squared value we can see that 46.37 percent of the variation is explained by the model. This is considerable lower then

the first model, and is only close to the first model with the points removed. This is one of the main reasons for opting for the first model. However, people may prefer this model as they might find it to be more intuitive then the previous model as annual holdings turnover is replaced by standard deviation. Standard deviation is a concept most people, especially in the financial world, have a grasp on. They understand it is a measure of the risk and can compare it easily to the return. For example if you tell some one in the financial sector a project has a standard deviation of 30%, they would expect the project to be yielding a high expected return to make up for the risk. Annual holdings turnover does not have that intuitive appeal. It's much more difficult to get a grasp on. If a company has a high holdings turnover you don't know if it is a risky company looking to capture the most out of a market, or a risk adverse company just looking to rebalance their portfolio to actually reduce the risk. This information on the goal of the mutual fund can easily be found in their prospectus; however it involves digging deeper whereas standard deviation does not.

Section 4. Summary and concluding remarks

Financial markets are some of the most complex and unpredictable things in the world. However, we have developed a model that is at least somewhat accurate in predicting year to date returns for mutual funds. The most accurate model we found used annual holdings turnover, worst 3-year return, and fund valuation as predictors.

The funds in this study were selected randomly from an online finance site, and hopefully represent an accurate cross section of the population. The data would certainly be more reliable if a larger sample had been taken. Also, there could be other methods of selection which guarantee a more random spread of funds. It is possible that this model does not predict well at very large and very small returns because the majority of funds chosen fell in generally the middle area. Perhaps there is a way to include high performance and low performance funds and still have a viable model. Also, it may be possible that there are other predictors that were overlooked or unavailable at the time of variable selection. Future studies could include some of these variables, or widen the sample selection.

Another way to possibly study this data would be as a time series model. In this way, fund performance for past years could be compared and used as a predictor for current success or failure. It would be interesting to see if certain firms have positive or negative trends, or even seasonal patterns. A time series would provide a new set of results that would be interesting to compare to the results from this study. However, this task is left for future researchers.

Appendix

APPENDIX TABLE OF CONTENTS:

1. References
2. Variable Definitions
3. Basic Summary Statistics
4. Initial Model and Diagnostics
5. Model and Diagnostics with Additional Variables

6. Checking for Unusual Observations
7. Checking for Variable Transformation
8. Model and Diagnostics with Categorical Variables Added
9. Alternative Model with Diagnostics

A.1 References

Faraway, Julian J. Linear Models with R. New York, Chapman and Hall/CRC, 2005.

http://finance.yahoo.com/, December 3, 2007.

http://www.smith.umd.edu/smithbusiness/spring2005/coverstory.html, 2007.

http://www.fool.com/mutualfunds/mutualfunds01.htm, 2007.

http://www.sec.gov/answers/mutfund.htm, 2007.

http://www.ici.org/about_ici.html, 2007.

Jensen, Michael C. "The Performance of Mutual Funds in the Period 1945 - 1964". University of Rochester College of Business.

A.2 Variable definitions

Variable	Definitions
Fnd.Sym	Fund ticker symbol
YTD.rtn	Year to date return (as of Dec. 3)
adj.2.yr.rtn	Adjusted two year return (2005-2006)
adj.4.yr.rtn	Adjusted four year return (2003-2006)
best.3.yr.rtn	Best three year return over the life of the fund
worst.3.yr.rtn	Worst three year return over the life of the fund
NA	Net Assets (in millions)
ER	Expense ratio
AHT	Annual holdings turnover
size	Size of funds investments, based on median market capitalization
valuation	Fund investment strategy (Value, Blend, or Growth)
SD	Standard deviation of fund returns
life	Lifetime of fund (measured in year)
stock	Percent portfolio composition attributed to stocks
cash	Percent portfolio composition attributed to cash
MSR	Morning star rating (out of 4 stars)
X3.yr.load.adj.rtn	Three year load adjusted return (same as three year return if no load)

A.3 Basic summary statistics

```
data<-read.csv("/Users/Miles/Desktop/YTD_Data.csv")
attach(data)
summary(data)
```

```
> summary(data)
    Fnd.Sym      YTD.rtn         adj.2.yr.rtn    adj.4.yr.rtn   best.3.yr.rtn
 AMGCX  : 1   Min.   :-1.610   Min.   :10.16   Min.   :10.01   Min.   : 8.54
 AMRGX  : 1   1st Qu.: 7.835   1st Qu.:12.34   1st Qu.:12.75   1st Qu.:15.38
 ATAFX  : 1   Median :11.075   Median :13.88   Median :15.36   Median :23.91
 AUXFX  : 1   Mean   :13.407   Mean   :14.85   Mean   :15.97   Mean   :22.94
 AVEGX  : 1   3rd Qu.:15.800   3rd Qu.:16.91   3rd Qu.:18.46   3rd Qu.:28.59
 BFOCX  : 1   Max.   :45.400   Max.   :22.51   Max.   :25.77   Max.   :60.82
 (Other):32

 worst.3.yr.rtn         NA.                  ER             AHT             size
 Min.   :-54.280   Min.   :    2.61   Min.   :0.000   Min.   :  0.00   Large :24
 1st Qu.:-10.047   1st Qu.:   21.23   1st Qu.:1.192   1st Qu.: 19.75   Medium: 6
 Median : -2.475   Median :   78.68   Median :1.390   Median : 51.00   Small : 8
 Mean   : -4.703   Mean   :  540.45   Mean   :1.467   Mean   : 73.37
 3rd Qu.:  3.603   3rd Qu.:  159.68   3rd Qu.:1.688   3rd Qu.:100.75
 Max.   : 16.540   Max.   :10370.00   Max.   :3.360   Max.   :386.00

  valuation        SD              life           stock            cash
 Blend :11   Min.   : 4.660   Min.   : 3.00   Min.   : 68.67   Min.   : 0.0000
 Growth:17   1st Qu.: 7.293   1st Qu.: 8.00   1st Qu.: 92.94   1st Qu.: 0.7975
 Value :10   Median : 9.560   Median :10.00   Median : 96.05   Median : 3.1900
             Mean   :10.155   Mean   :13.26   Mean   : 94.23   Mean   : 4.3547
             3rd Qu.:12.398   3rd Qu.:14.00   3rd Qu.: 98.67   3rd Qu.: 5.9925
             Max.   :22.210   Max.   :61.00   Max.   :100.00   Max.   :14.9900

      MSR        X3.yr.load.adj..rtn
 Min.   :1.000   Min.   : 8.59
 1st Qu.:3.000   1st Qu.:11.13
 Median :3.000   Median :12.59
 Mean   :3.026   Mean   :13.70
 3rd Qu.:4.000   3rd Qu.:15.82
 Max.   :4.000   Max.   :25.52
```

For a response variable we choose Year to Date Returns on the mutual funds. We had a total of 38 observations of mutual funds. The data was retrieved from November 30th 2007 from Yahoo Financial. The mutual funds chosen were stock mutual funds as opposed to Bond or municipal mutual funds that have different factors affecting them such as their tax benefits. The returns on the mutual funds have a wide spread ranging from a negative value all the way up to a 45% return. Looking at the data there appears to be some linear relationship between YTD returns and 2 year adjusted return, 5 year adjusted return, Size, Standard Deviation, and Annual holdings turnover.

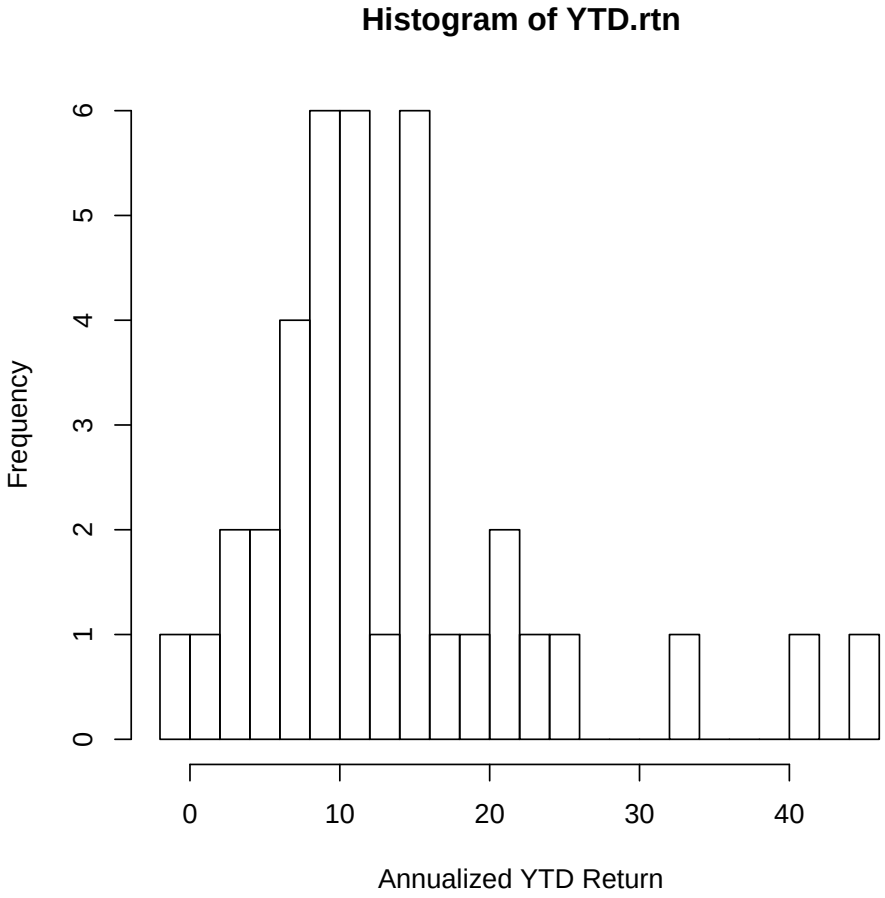

Figure 1: Histogram of Year to Date Returns appears to be somewhat skewed, however is relatively normal. The graph (not shown) of the log of YTD Returns is a little more normal shaped however it does not work for the negative return values and its interpretation is not intuitive so we opted not to use the log transformation on the dependent variable.

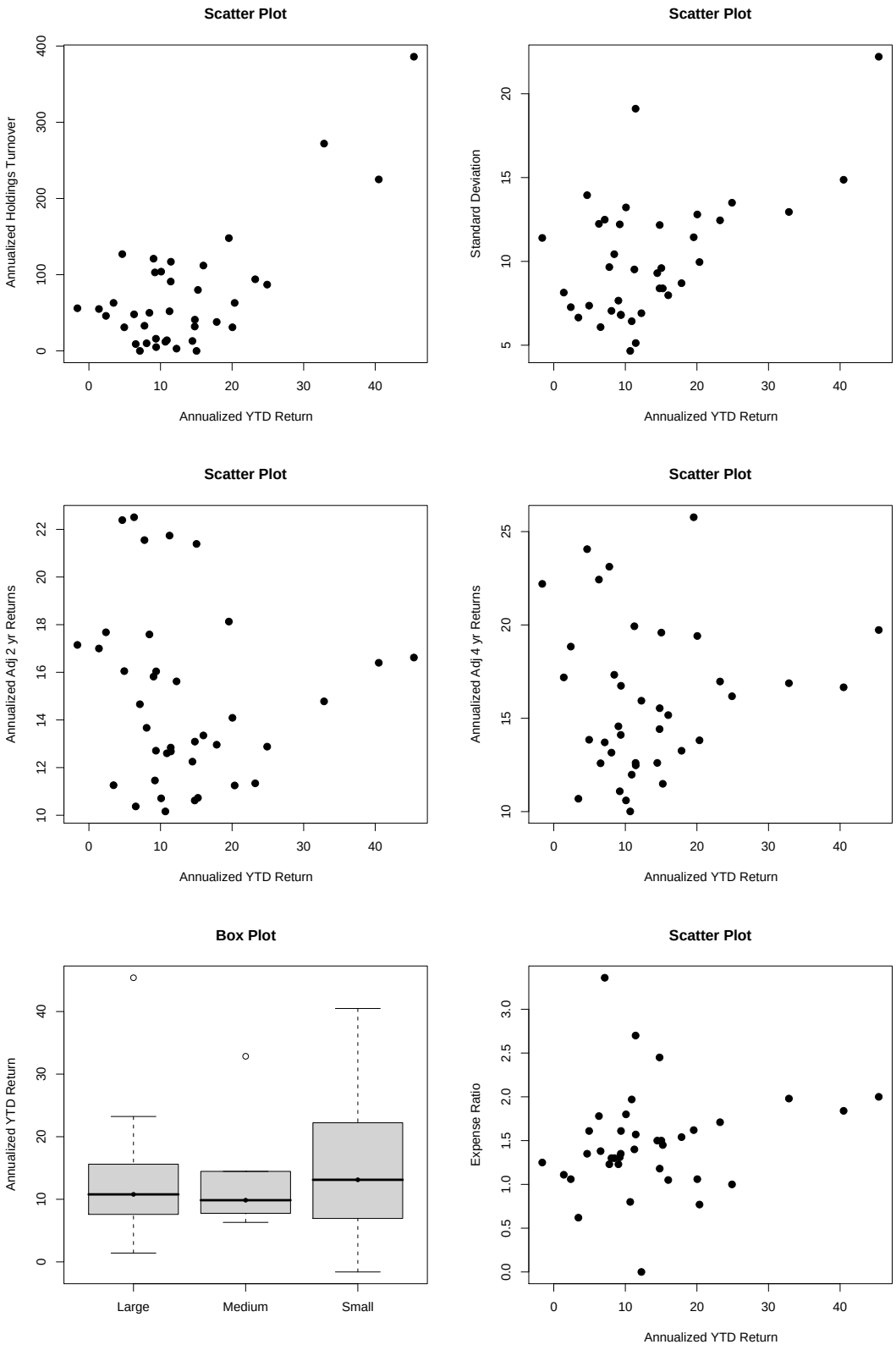

Figure 2: A look at the scatter plots of the variables that appeared correlated in the data.

From the histograms and box plots the best linear explanatory variables seem to be standard deviation, annual holdings turnover, and the size of the companies in which it invests, especially in the medium case.

```
> cor(YTD.rtn,AHT)
[1] 0.7276816

> cor(YTD.rtn,SD)
[1] 0.5601571
```

Because both annual holdings turnover and standard deviation looked like good variables to start the model, we looked at the correlation coefficients of these variables and determined that annual holdings turnover had a better fit.

A.4 Initial model and diagnostics

As a result of our preliminary data analysis, we choose annual holdings turnover to be our first explanatory variable:

```
> lm1<-lm(YTD.rtn~AHT)
> summary(lm1)

Call:
lm(formula = YTD.rtn ~ AHT)

Residuals:
     Min       1Q   Median       3Q      Max
-13.6047  -4.7566   0.3758   4.9214  13.3487

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.76130    1.52578   4.431 8.41e-05 ***
AHT          0.09058    0.01423   6.365 2.26e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.859 on 36 degrees of freedom
Multiple R-squared: 0.5295,Adjusted R-squared: 0.5165
F-statistic: 40.52 on 1 and 36 DF,  p-value: 2.264e-07
```

This first model appears to be decent. We can see that the variable annual holdings turnover is significant as the p-value is 2.26e-07 which is significant to almost any alpha level. Also the model has an adjusted R-squared value of .5165. In other words 51.65% of variation is explained by our model, which is excellent considering we have only used one variable.

The next step is to run diagnostics on the model to make sure the residuals are normal and have constant variance. Also it is important to look for outliers and leverage points.

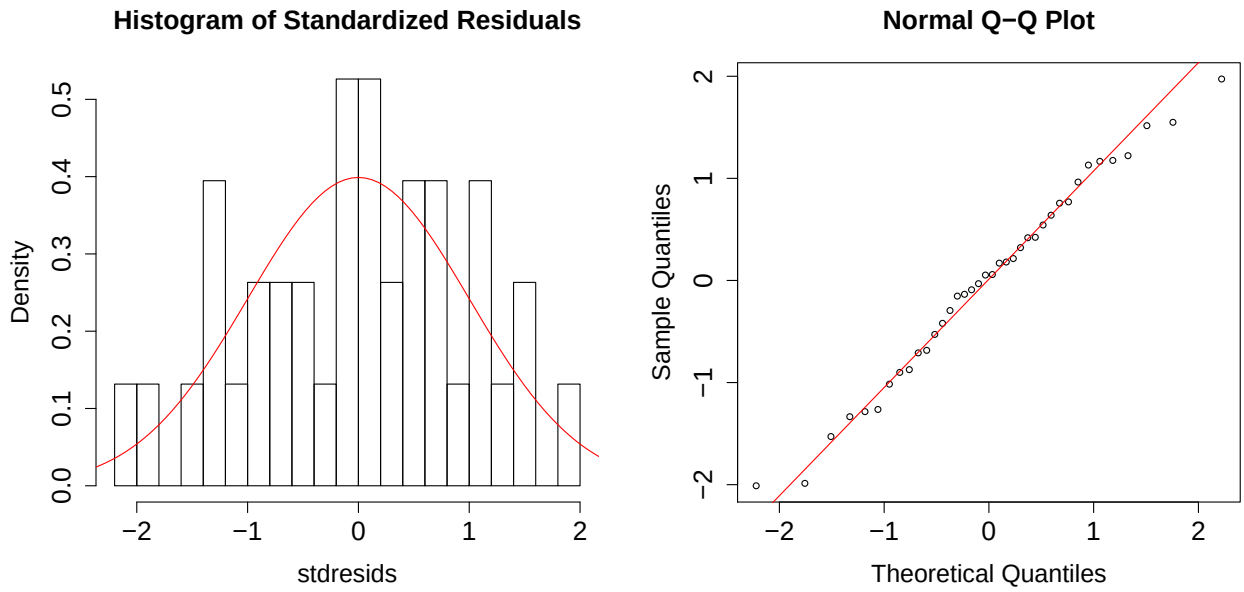

Figure 3: Histogram and qq plots of the standardized residuals.

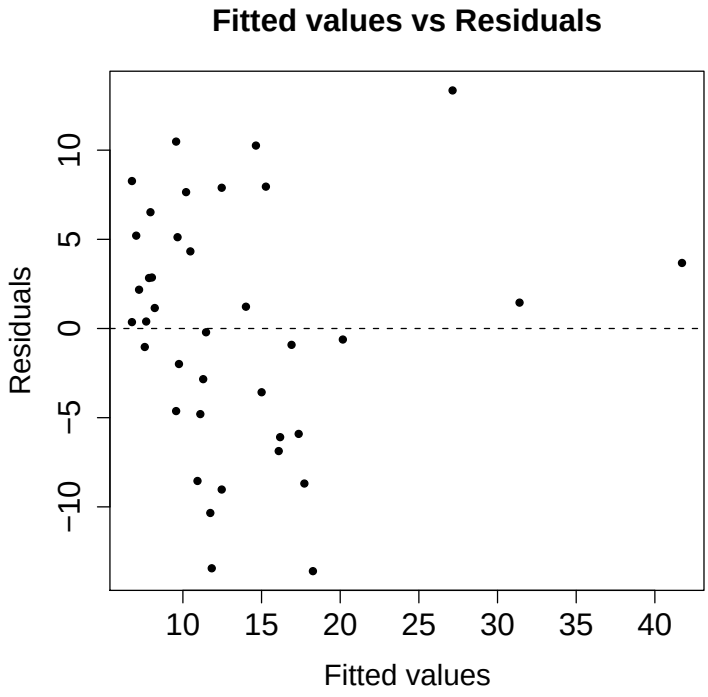

Figure 4: Examining constant variance - fitted values vs residuals.

The histogram of the residuals (Figure 3) does not look as normal as we would hope, however this is probably due to our original year to date returns being somewhat skewed. However the qq plot tells a different story. The qq plot shows our residuals following very closely to a normal distribution, even in the tails.

Looking at the fitted vs. residuals plot (Figure 4) it appears that the residuals has constant variance which we were hoping to find. It also appears however that there is limited data for YTD returns greater than 25 percent. Next we will look for outliers or leverage points that could be affecting our model.

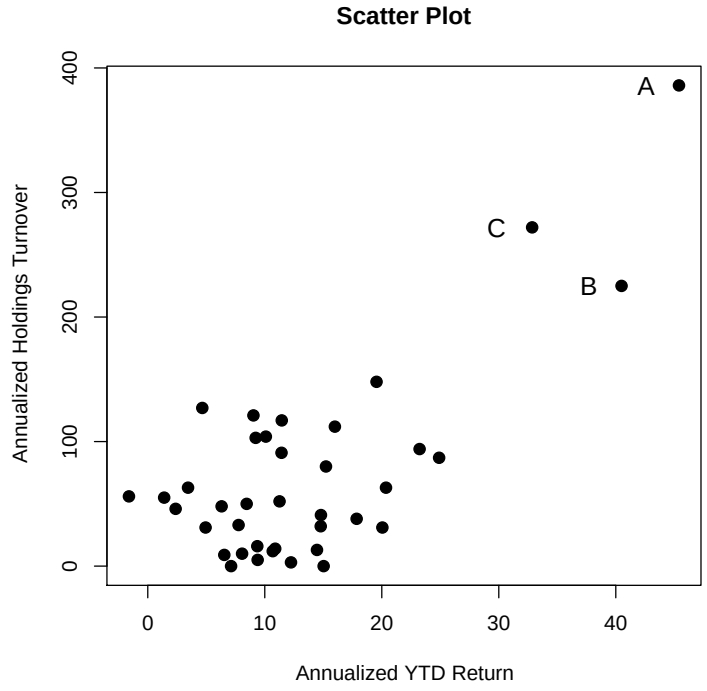

Figure 5: Examining potential outliers.

Looking at the original scatter plot (Figure 5) of annual holdings turnover and year to date returns there looks as thought the maybe be 3 potential outliers. Point A (pt 5) within the data which has an AHT of 386 and YTD return of 45.40% is the furthest from the data. The other points that appear to be leverage points are at B (pt 15) and C (pt 31). They have YTD returns of 40.49% and 32.85% respectively, and annual holdings turnover of 225 and 272 again respectively. The table below shows the effects on the model of removing one, two or all three of the points in comparison to the base model.

Data	b0	b1	R-squared(%)	p-value(b1)
Original	6.76130	0.09058	0.5295	2.26e-07
Without A	7.24258	0.08163	0.345	0.000132
Without B	7.09043	0.08062	0.4766	2.27e-06
Without A and B	8.23097	0.05845	0.2067	0.00534
Without A, B, and C	9.71449	0.02662	0.03278	0.298

Clearly removing the outliers has a drastic effect on the linear model. This shows the great influence of these points. In seeing the great impact these points have on the model I decided to look further into them. Point A is mutual fund Berkshire Focus. Its return of 45.40% is not all that surprising as it has had a 1 year best total return of 142.90% and a 3 year best total return of 60.82%. This mutual fund has had very big returns in the past but it also has a high standard deviation of 32.21 over the last 5 years. Also a possible reason for its high year to date return could be due to the returns of some of its major investments. It has 12.17% of its investments in the company Research in Motion which has a Year to date return of 192.32%. In looking at point B, Lord Abbett Developing Growth C Mutual Fund, we see that it is a little more unusual. This fund is currently at a record high as its previous 1 year best was 39.19% compared to its year to date return of 40.49%. However, this is not extremely unusual as it is very close to its previous high. Its high return is not accounted for by either its standard deviation, which is relatively low at 14.87 over the last 3 years, or its 3 year average return, which is 23.93%. In looking at the third point, Alger MidCap Growth C, it is similar to the first outlier as it has had a 1 year high in the past of 44.36% and a 1 year worst of -30.82% with a standard deviation of 14.55 of the last 5 years. Based on this information it would be most logical to remove data point B, or mutual fund LADCX. However we will leave all of the data points in the set while we add other variables. If at that time it is still an outlier we will contemplate removing it then.

A.5 Model and diagnostics with additional variables

The next step is to look for other variables that may be useful in the model. We added each of the other non-categorical variables individually and looked at the p-value of that variable and the adjusted R-squared value of the model and determined that the only three variables to have a p-value less then 15% were the returns over the last two years, how long the mutual fund has been around and the worst 3 year returns of the fund over its life. We decided to work with the worst 3 year returns over the life of the fund as it had the only significant p-value, as well as the best adjusted R-squared. We also opted out of using the log transformation on this variable because it is a percentage and is much more intuitive without the transformation. Also we found the p-value to be lower without the log transformation.

```
> cor(cbind(YTD.rtn,AHT,worst.3.yr.rtn))
                  YTD.rtn        AHT worst.3.yr.rtn
YTD.rtn         1.0000000  0.7276816     -0.6078166
AHT             0.7276816  1.0000000     -0.4124881
worst.3.yr.rtn -0.6078166 -0.4124881      1.0000000
```

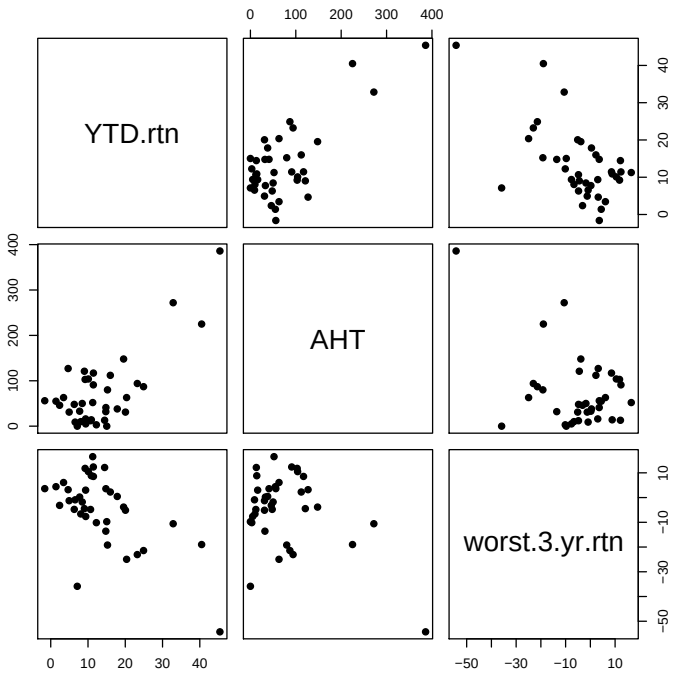

Figure 6: This shows some of the correlation between the different variables and the main dependent variable. Looking at this there doesn't appear to be a great correlation between annual holdings turnover and the 3 year worst returns, which is good as they are both explanatory variables. Unfortunately there doesn't appear to be a great amount of correlation between the year to date returns and the 3 year worst returns, but the adjusted R-squared value suggests that there is enough.

```
Summary statistics with the new added variable
> lm2<-lm(YTD.rtn~AHT + worst.3.yr.rtn)
> summary(lm2)
Call:
lm(formula = YTD.rtn ~ AHT + worst.3.yr.rtn)
Residuals:
     Min       1Q   Median       3Q      Max
-11.6440  -2.9424   0.5709   3.7065  12.5742
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.95264    1.34806   5.158 9.96e-06 ***
AHT             0.07154    0.01379   5.188 9.07e-06 ***
worst.3.yr.rtn -0.25625    0.07657  -3.347  0.00196 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 6.055 on 35 degrees of freeom
Multiple R-squared: 0.6436,Adjusted R-squared: 0.6232
F-statistic:  31.6 on 2 and 35 DF,  p-value: 1.443e-08
```

Before we get too far ahead of ourselves let's interpret what we have found so far.

Expected YTD.rtn = 6.95264 + AHT*.07154 + worst.3.yr.rtn*-.25625

This model is predicting the year to date return of a mutual fund based on the mutual funds annual holdings turnover and its worst 3-year returns. The intercept suggests that a mutual fund with no holdings turnover, or one that keeps the same holdings the entire year, and has had a zero percent return as its worst 3-year return will return about 6.95% to date. This seems to make sense, as it is slightly higher than the return on government bonds and it is expected that a mutual fund with all its expertise could choose a set of holdings at the beginning of the year that would do better then the risk free rate (government bonds). The annual holdings turnover coefficient of .071 suggests that with all else being held constant if a mutual fund increase its annual holdings turnover by 1 percent its expected year to date return should increase by .07154 percent. The worst 3 year return coefficient suggests that if a mutual funds worst 3 year return were 1 percent higher it would expect its year to date return to decrease by .25625%. This doesn't seem to make sense at first glance, but if you think about it for a second it does hold water. The lower the worst 3 year return would suggest a higher standard deviation on the mutual fund, and a high standard deviation leads to the potential for greater returns in this year.

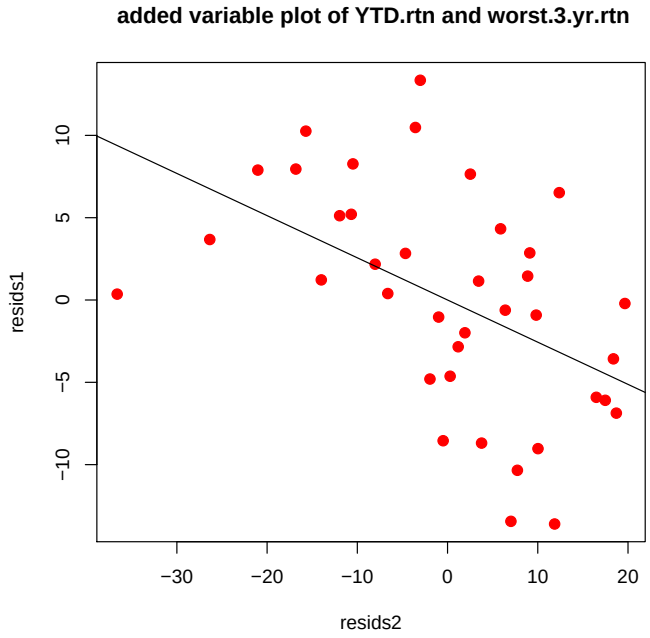

Figure 7: This is an added variable plot which allows us to look at the effect of the worst 3-year return on the year to date return while removing the effect of the annual holdings turnover. In this we look for nonlinearity and outliers and/or influential points. From what we can see there doesn't seem to be a great deal of linearity, although some. As well there doesn't appear to be any influential points or outliers.

```
> cor(resids1,resids2)
[1] -0.4923741
```

Next we must do further diagnostics of the model:

```
> lm3<-lm(YTD.rtn~I(AHT + worst.3.yr.rtn))
> anova(lm3,lm2)
Analysis of Variance Table
Model 1: YTD.rtn ~ I(AHT + worst.3.yr.rtn)
Model 2: YTD.rtn ~ AHT + worst.3.yr.rtn
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     36 2043.17
2     35 1283.03  1    760.14 20.736 6.128e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

To test that the coefficients of the two variables are in fact different, we test a subspace of the model using an indicator variable. From the p-value of 6.128e-05 we can see there is clearly no evidence that the coefficients between annual holdings turnover and worst 3-year return are equal.

```
> confint(lm2)
                    2.5 %      97.5 %
(Intercept)     4.21592834  9.68935761
AHT             0.04354984  0.09953533
worst.3.yr.rtn -0.41168413 -0.10080779
```

This again is further evidence that all the variables are significant as none of the confidence intervals contain zero. Next we need to check that the errors of the new model are independent, have constant variance and are normally distributed.

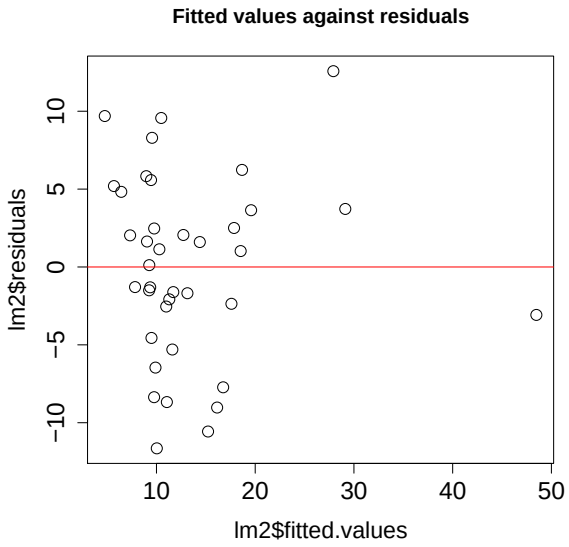

Figure 8: Again we see similar results as we did with the simple linear model. The model has constant variance in the lower year to date returns but has limited coverage over very high year to date returns.

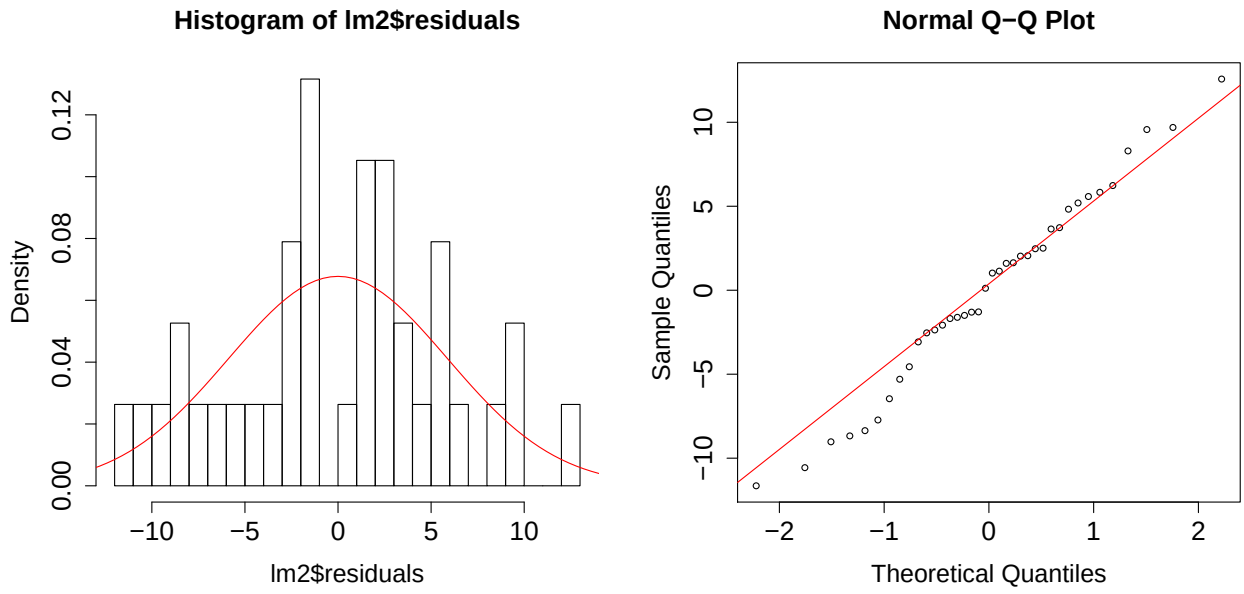

Figure 9: From the histogram and qq plot we can see that the residuals are somewhat normal shaped. The histogram is a little fat in the tails, but is peaked right around zero. The qq plot shows some separation from the normal line, especially left bound.

```
> library(lmtest)
Loading required package: zoo
> dwtest(lm2)

Durbin-Watson test

data:  lm2
DW = 2.6145, p-value = 0.9775
alternative hypothesis: true autocorrelation is greater than 0
```

From the Durbin-Watson test yielding a high p-value of .9775 we can see that we fail to reject the null hypothesis that there is no correlation. This is good as it proves that the residuals are uncorrelated.

A.6 Checking for unusual observations

Next we look for outliers and/or leverage points:

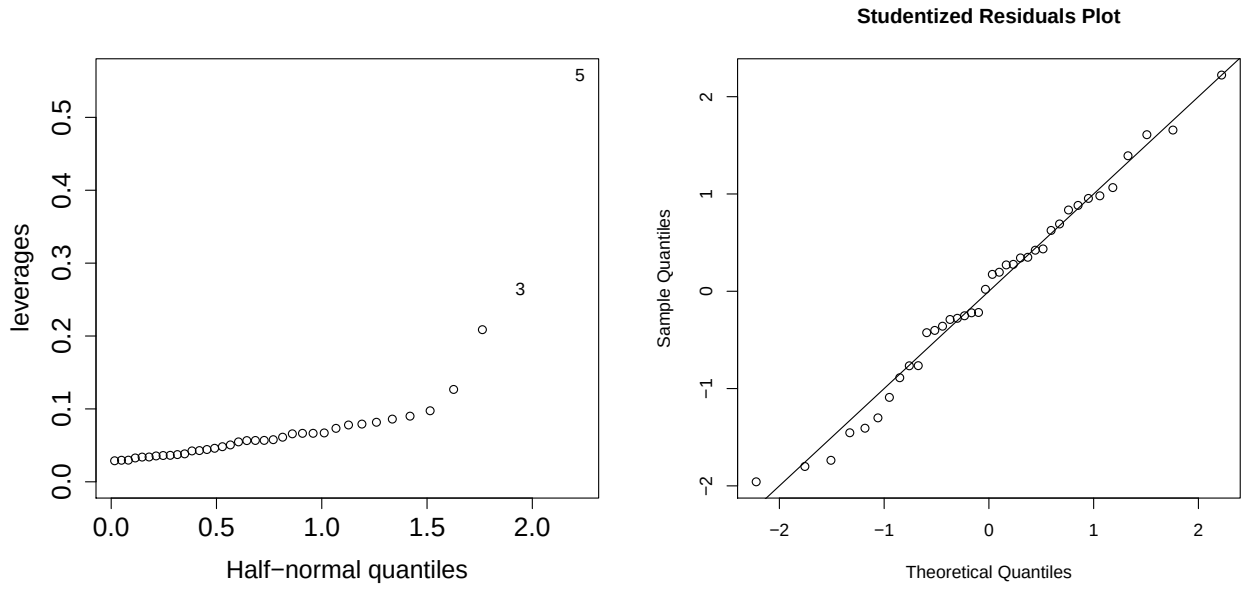

Figure 10: From the half-normal plot it appears as though points 3 and 5 are leverage points. However, from the studentized residuals plot there doesn't appear to be any leverage points.

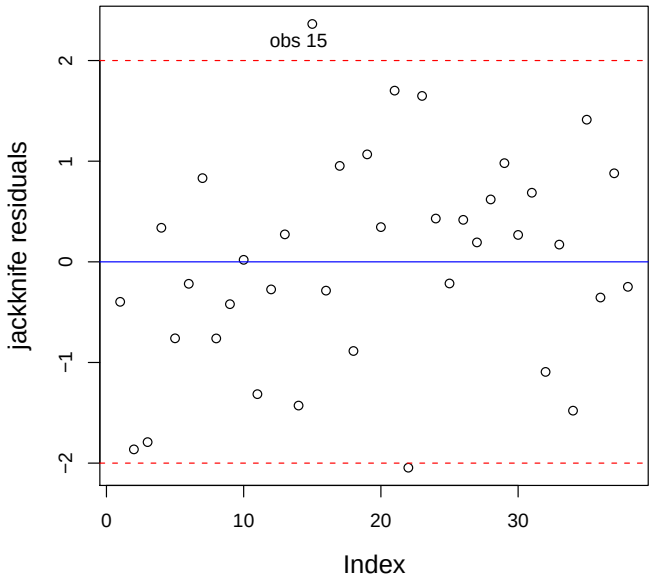

Figure 11: From the graph of the jackknife residuals there appears to be only 2 possible outliers. The one at the bottom of the graph is very close to the line y=-2 which is our rule of thumb for declaring outliers with jackknife residuals. The only point that appears to be a true outlier is point 15 at the top of the graph. This is the same point that we believed to be an outlier in the simple linear regression.

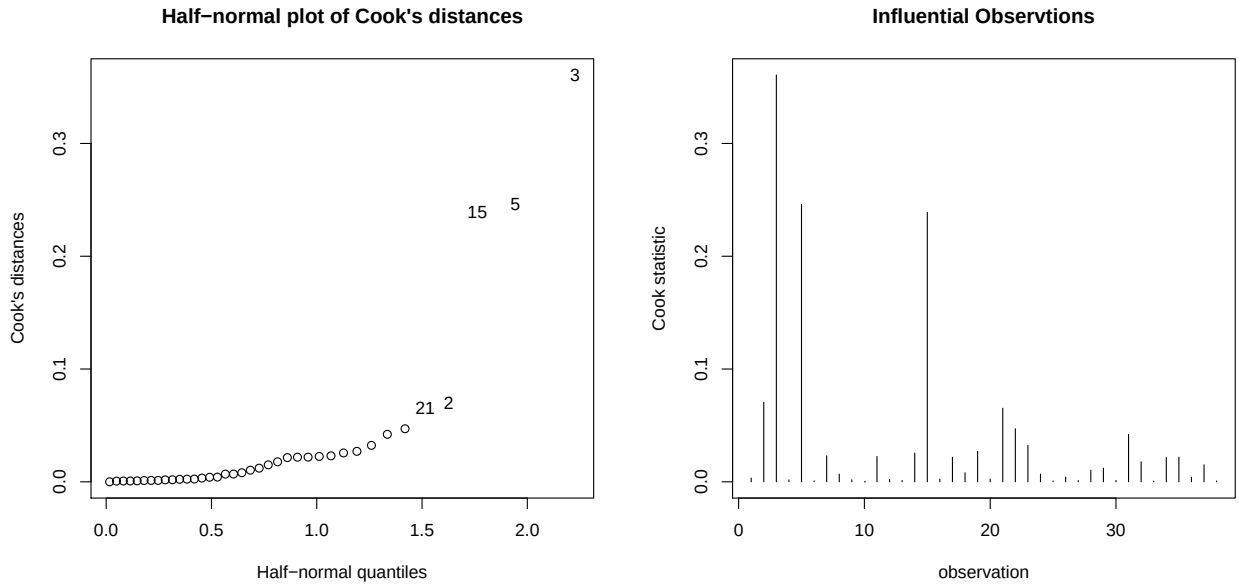

Figure 12: On the left is a half normal plot of Cook's distances. On the right is a time series graph using the Cook statistic. Both of these are helpful in detecting influential observations. From the graphs it appears the points 3, 5 and 15 are influential points.

Below is the summary of the model without point 3, then without point 5, and then without point 15 as well, since they seem to be the most influential on the model.

Without point 3

```
Call:
lm(formula = YTD.rtn ~ AHT + worst.3.yr.rtn, subset = (Fnd.Sym !=
    "AMRGX"))

Residuals:
    Min      1Q  Median      3Q     Max
-11.530  -3.665   0.632   4.222  12.622

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.61291    1.35842   5.604 2.82e-06 ***
AHT             0.06234    0.01433   4.351 0.000117 ***
worst.3.yr.rtn -0.32804    0.08438  -3.887 0.000447 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.872 on 34 degrees of freedom
Multiple R-Squared: 0.6706,Adjusted R-squared: 0.6512
F-statistic: 34.61 on 2 and 34 DF,  p-value: 6.329e-09
```

Without point 5

```
Call:
lm(formula = YTD.rtn ~ AHT + worst.3.yr.rtn, subset = (Fnd.Sym !=
    "BFOCX"))

Residuals:
     Min       1Q   Median       3Q      Max
-11.4584  -3.0194   0.3412   2.7760  10.8816

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.47039    1.49742   4.321 0.000128 ***
AHT             0.07873    0.01679   4.689 4.34e-05 ***
worst.3.yr.rtn -0.28559    0.08617  -3.314 0.002190 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.091 on 34 degrees of freedom
Multiple R-Squared: 0.505,Adjusted R-squared: 0.4759
F-statistic: 17.34 on 2 and 34 DF,  p-value: 6.438e-06
```

Without point 15

```
Call:
lm(formula = YTD.rtn ~ AHT + worst.3.yr.rtn, subset = (Fnd.Sym !=
    "LADCX"))

Residuals:
     Min       1Q   Median       3Q      Max
-11.4772  -2.3873   0.4232   4.3357   9.5731

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.25800    1.27414   5.696 2.14e-06 ***
AHT             0.06266    0.01350   4.642 4.98e-05 ***
worst.3.yr.rtn -0.24929    0.07205  -3.460  0.00148 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.693 on 34 degrees of freedom
Multiple R-Squared: 0.6129,Adjusted R-squared: 0.5901
F-statistic: 26.91 on 2 and 34 DF,  p-value: 9.857e-08
```

Without all three points

```
Call:
lm(formula = YTD.rtn ~ AHT + worst.3.yr.rtn)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3231  -3.0128   0.6196   3.3747   9.9333

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.81476    1.44081   5.424 5.79e-06 ***
AHT             0.05513    0.01744   3.162  0.00342 **
worst.3.yr.rtn -0.32933    0.09218  -3.573  0.00114 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.56 on 32 degrees of freedom
Multiple R-Squared: 0.4261,Adjusted R-squared: 0.3903
F-statistic: 11.88 on 2 and 32 DF,  p-value: 0.0001383
```

Clearly these data points have an effect on the model, however all the variables are still significant and the model still explains 39.03% of the variation even without all three, which is not all that terrible.

A.7 Checking for variable transformations

Next we checked to see if any of the variables might fit better with a quadratic relationship. To check this we used a scatter plot of the residuals with loess to smooth the curve. This is done on the annual holdings turnover to the right. We noticed that this appeared to have some sort of a quadratic shape, so we tried fitting the model with the annual holdings turnover squared term. Below is a summary that showed that. We found that the explanatory variables were no longer significant with p-values greater then 5%. Because of this we opted to stay with our original model.

```
Call:
lm(formula = YTD.rtn ~ AHT + AHT.sq)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8500  -4.2175  -0.1141   3.9168  15.1970


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.1993604  1.9419202   4.737 3.54e-05 ***
AHT         0.0248132  0.0368318   0.674   0.5049
```

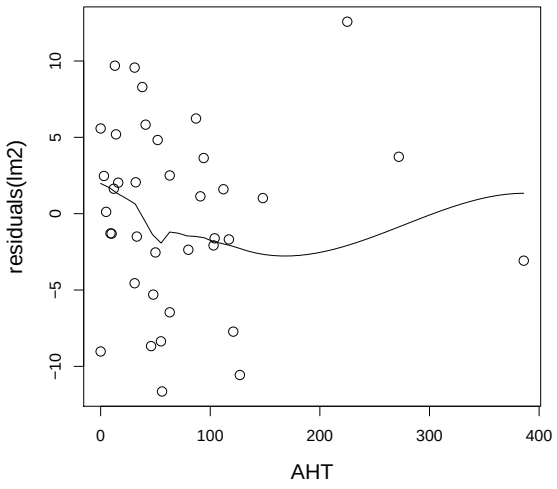

Figure 13: Examining possible transformation of variable AHT.

```
AHT.sq      0.0002076  0.0001079   1.924   0.0625 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.615 on 35 degrees of freedom
Multiple R-Squared: 0.5745,Adjusted R-squared: 0.5502
F-statistic: 23.63 on 2 and 35 DF,  p-value: 3.201e-07
```

A.8 Model and diagnostics with categorical variables added

The next step is to look at the effects of adding one of the categorical variables to the model. The valuation variable was the only categorical variables that helped the model at all. In looking at the size variable we found the adjusted R-squared value decreased. The valuation variable looks at whether the mutual fund is a growth mutual fund, a value mutual fund or a blend of the two. The summary with that added variable is below:

```
Call:
lm(formula = YTD.rtn ~ AHT + worst.3.yr.rtn + valuation)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2638  -4.3969   0.8427   3.8713  11.6020

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.50588    1.86028   4.035 0.000305 ***
AHT             0.07146    0.01377   5.191 1.05e-05 ***
```

```
worst.3.yr.rtn  -0.20680    0.07930  -2.608 0.013582 *
valuationGrowth  1.37619    2.45315   0.561 0.578595
valuationValue  -3.53535    2.61772  -1.351 0.186031
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.899 on 33 degrees of freedom
Multiple R-Squared: 0.681,Adjusted R-squared: 0.6424
F-statistic: 17.62 on 4 and 33 DF,  p-value: 7.93e-08
```

From this we see that the r-squared value increased slightly from 62.32% to 64.24%. In looking at the p-values it appears that none of the valuation terms are significant, however the intercept is extremely significant and that holds the information of the Blend valuation. Now the model takes the shape of:

> Predicted YTD rtn = 7.50588 + .07146*AHT + -.2068*worst 3 year rtn + 1.37619*valuationGrowth + -3.53535*valuationValue

Because there are the categorical terms in the model it works in the following way. If the mutual fund is a Blend mutual fund, then the valuationGrowth and valuationValue terms are zero and the model looks like the following

> Predicted YTD rtn = 7.50588 + .07146*AHT + -.2068*worst 3 year rtn

If the mutual fund of the type Value then the valuationValue term is 1 and the valuationGrowth term is zero. The model then looks like

> Predicted YTD rtn = 7.50588 + .07146*AHT + -.2068*worst 3 year rtn + -3.53535*valuationValue

If the mutual fund is of the type Growth then the valuationValue term is 0 and the valuationGrowth term is 1. The model then looks as such

> Predicted YTD rtn = 7.50588 + .07146*AHT + -.2068*worst 3 year rtn + 1.37619*valuationGrowth

The interpretation for the coefficients is the same for the annual holdings turnover and the worst 3-yr returns. The intercept is a little different as it now accounts for the valuation-Blend term as well. It now suggests that a blend type mutual fund with zero annual holdings turnover and a 3-yr worst return of zero has an expected year to date return of 7.50588%. The coefficient for valuationValue is the expected difference in the year to date return for a Blend mutual fund and a Value mutual fund. The coefficient for valuationGrowth is very similar except it is the expected difference in the year to date return for a Blend mutual fund verses a Growth mutual fund.

Next we will run some diagnostics on the new model and check to see if there is any collinearity between the variables. We will do so using the VIF, or the variance inflation factor.

```
> vif(z)
            AHT  worst.3.yr.rtn valuationGrowth  valuationValue
       1.265293        1.362004        1.624951        1.451204
```

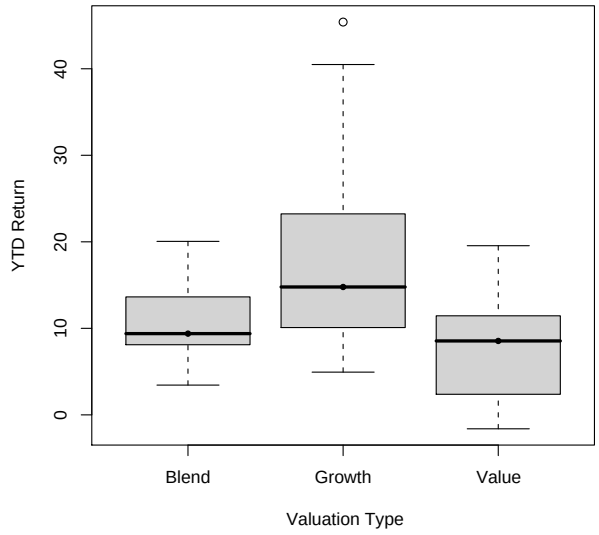

Figure 14: Here is a boxplot of the categorical variable valuation to help in visualizing the data. There seems to be some differences in the returns based on the valuation, however it is not huge.

From this we can see that there is not severe collinearity. As a rule of thumb severe collinearity is when the VIF exceeds 10.

```
Call:
lm(formula = YTD.rtn ~ valuation)

Residuals:
    Min      1Q  Median      3Q     Max
-13.256  -6.575  -1.285   4.068  27.204

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       10.660      2.728   3.908 0.000407 ***
valuationGrowth    7.536      3.501   2.153 0.038303 *
valuationValue    -2.374      3.953  -0.601 0.551987
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.047 on 35 degrees of freedom
Multiple R-Squared: 0.2042,Adjusted R-squared: 0.1588
F-statistic: 4.492 on 2 and 35 DF,  p-value: 0.01835

> anova(lmtest)
Analysis of Variance Table

Response: YTD.rtn
```

```
           Df  Sum Sq Mean Sq F value  Pr(>F)
valuation   2  735.22  367.61  4.4916 0.01835 *
Residuals  35 2864.55   81.84
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

In looking at both the anova table and the summary of a linear model of just valuation we can see that valuation is significant and that it has a pretty good R-squared value of .2042. Next we will look at the Tukey Confidence Intervals to see which of the valuation types are actually significant in comparisons to the others.

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = YTD.rtn ~ valuation)

$valuation
                  diff        lwr       upr     p adj
Growth-Blend   7.53647  -1.030661 16.103602 0.0938467
Value-Blend   -2.37400 -12.047653  7.299653 0.8207087
Value-Growth  -9.91047 -18.733843 -1.087098 0.0248530
```

In looking at this we can see that only Value and Growth mutual funds are significantly different, as zero is not within the confidence interval. Next we will do a little diagnostics on the new model with the categorical variable:

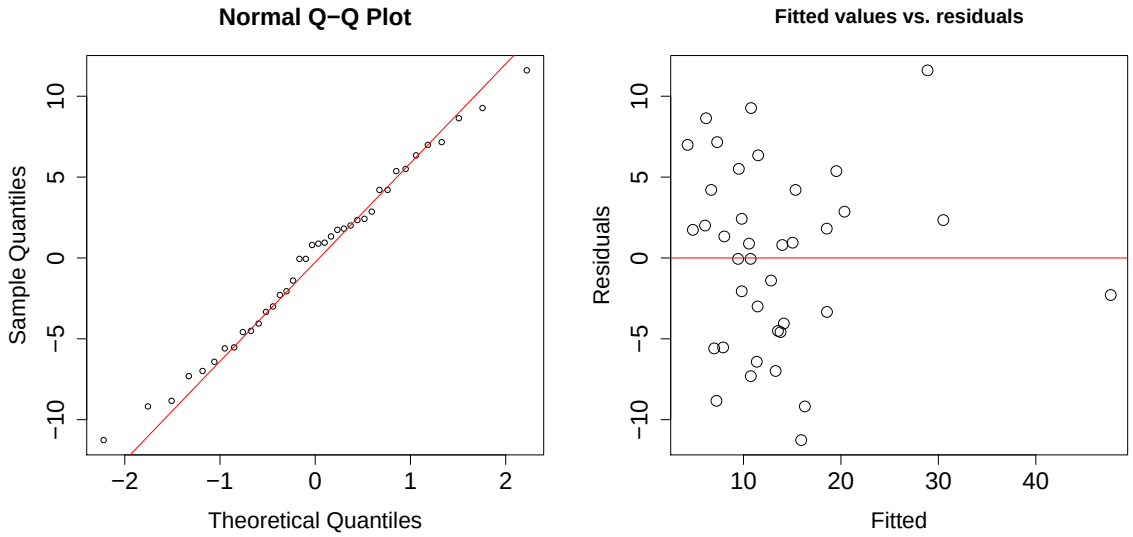

Figure 15: In looking at the Normal qq plot the residuals appear to be somewhat normal. They seem to fit the worst in the middle and the two tails. In the fitted vs. residuals plot we can see the variance is relatively constant, and again has the three points separated from the rest of the data.

A.9 Alternative model with diagnostics

After this we checked to see if any other variables could be added to the model and had no success as each of the additions were either insignificant or caused a another variable to become insignificant. Thus we decided to see if there were any other models that would be a good predictor of the year to date returns that people may like better. We found the following model to work as well as be intuitive:

Predicted YTD Returns = 1.58998 + 1.01659*SD + -.31745*worst 3-year return

```
Call:
lm(formula = YTD.rtn ~ SD + worst.3.yr.rtn)

Residuals:
    Min      1Q  Median      3Q     Max
-18.554  -3.430   0.375   3.956  17.755

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.58998    3.59380   0.442  0.66091
SD              1.01659    0.34862   2.916  0.00615 **
worst.3.yr.rtn -0.31745    0.09036  -3.513  0.00124 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.223 on 35 degrees of freedom
Multiple R-Squared: 0.4927,Adjusted R-squared: 0.4637
F-statistic:    17 on 2 and 35 DF,  p-value: 6.955e-06
```

From the summary we can see that both of the explanatory variables are significant to an alpha of .01, which is very strong. Also from the adjusted R-squared value we can see that 46.37% of the variation is explained by the model. This is not bad considering the difficulty of creating a model for mutual funds. We find this model to be more intuitive than the previous model as annual holdings turnover is replaced by standard deviation.

And finally, we ran a few diagnostics on the new model and some results shown in Figure 16.

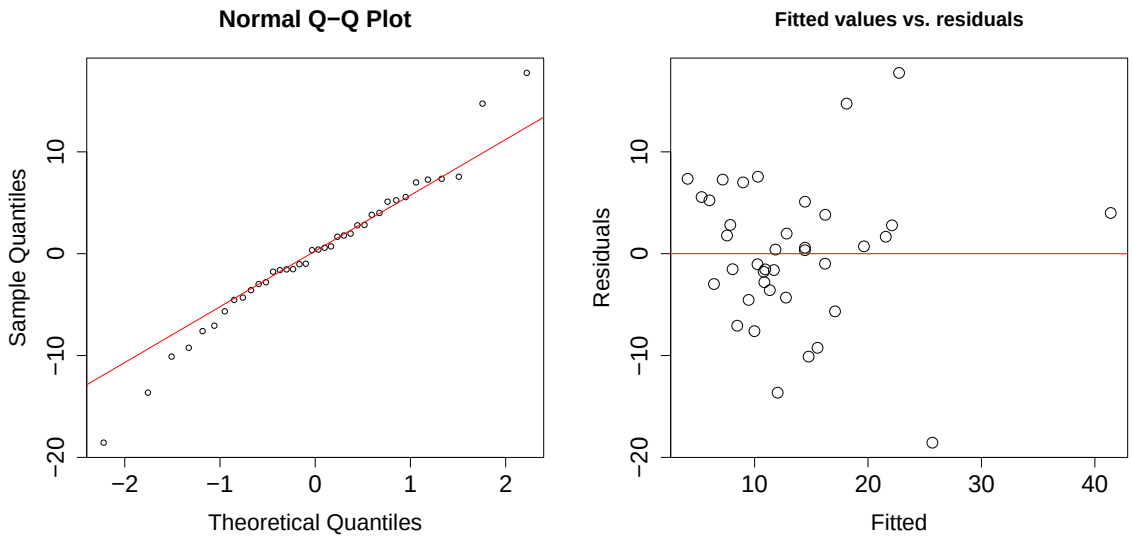

Figure 16: In looking at the normal qq plot we can see the residuals stray from normal in the tails but fit decently well in the middle of the data. Also the Fitted vs. Residuals plot appears to show the variance decreasing, but this could also be due to having such a small data set.