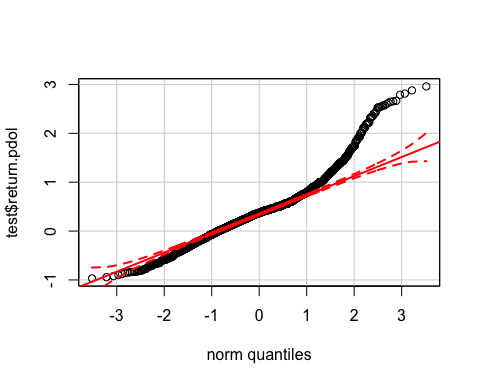
Variable Selection

Amy Cook

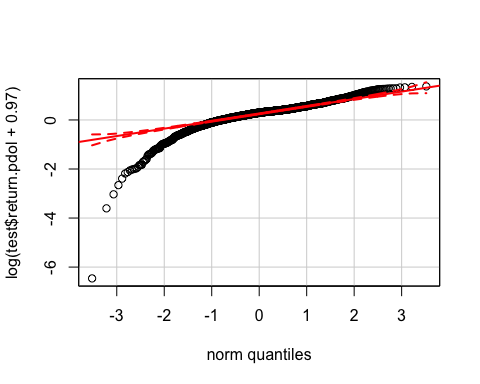
Thursday, June 18, 2015

Check for normality of return variable return per dollar Outliers were deleted first

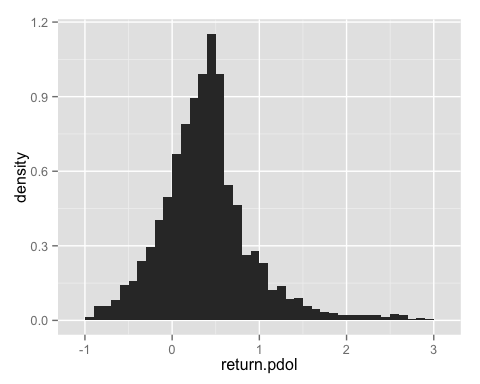
# try deleting very high return.pdol values greater than 3  
test<- all7 %>% filter(return.pdol <=3 & return.pdol> -2)  
qqPlot(test$return.pdol)



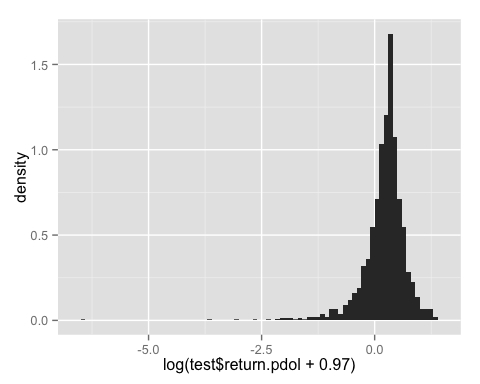
qqPlot(log(test$return.pdol+.97))



#have a look at the shape of return.pdol as bar graph  
ggplot(test, aes(x=return.pdol)) + geom\_histogram(aes(y=..density..), binwidth=.1)



ggplot(test, aes(x=log(test$return.pdol+.97))) + geom\_histogram(aes(y=..density..), binwidth=.1)



definietly don't want to use log - worse. Let's look at standard deviation, mean, median, skewness, kurtosis

sd(test$return.pdol)

## [1] 0.5310226

mean(test$return.pdol)

## [1] 0.3915029

median(test$return.pdol)

## [1] 0.3713105

#mode  
names(sort(-table(test$return.pdol)))[1]

## [1] "0.5"

skewness(test$return.pdol)

## [1] 0.9654046

skewness(log(test$return.pdol+.97))

## [1] -2.676149

kurtosis(test$return.pdol)

## [1] 2.789816

skewness only 0.92!! for unlogged skewness is -2.6 for logged data

kurtosis equals 2.85 - very peakedy - ie not flat. normal distribution is 0.

Create new data set that has outliers deleted, all7a

### delete outliers! and variables that aren't by milestone  
all7a<- all7 %>% filter(return.pdol <=3 & return.pdol> -2)  
all7a<- all7 %>% select(-Paperless,-Innovation, -JobInvCount,-job.first.inv.email.Sent,   
 -Total.Costs..AUD., -Stage, -Folders, -Total.Fee, -Disburse, -Subcon.fee,  
 -Job.Size, -Tot.Invoiced, -charge, -cost, -Dis.subcon, -hours,-profit,-balance,  
 -Role)

# Variable Selection

It would be good to be able to narrow down the variables to the correlated or strongly affecting ones. This improves accuracy and speed of the final machine learning algoriths.

Anova and Linear regression will give us an idea of variable importance. Look at coefficients

In order to get Anova and linear regression to work on this data set, start with the variables that are complete or almost complete. Then can add additional variables one by one. The function below does this automatically.

For lm, the output is ordered from abs(highest coefficient)

# BT.lin(type= 'lm', data= all7a[1:1313,], add.var = 'Job.Type.Primary') %>% slice(1:20)  
BT.lin(type= 'lm', data= all7a, add.var = 'Job.Type.Primary') %>% slice(1:20)

## Estimate Std. Error t value Pval  
## 1 60.0894393 82.3443829 0.7297333 4.656588e-01  
## 2 -2.5556214 0.4654543 -5.4905956 4.643490e-08  
## 3 -2.3041613 0.9688363 -2.3782772 1.750919e-02  
## 4 -2.2813506 1.1218475 -2.0335658 4.215819e-02  
## 5 -2.1978412 0.6008390 -3.6579535 2.623772e-04  
## 6 1.4729788 1.3800184 1.0673617 2.859677e-01  
## 7 -1.3308285 0.9241970 -1.4399837 1.500653e-01  
## 8 1.2283905 0.7198141 1.7065386 8.809961e-02  
## 9 -1.0483926 1.7409462 -0.6021970 5.471273e-01  
## 10 -0.9610273 0.7531682 -1.2759797 2.021457e-01  
## 11 -0.8643256 0.5103767 -1.6935053 9.055169e-02  
## 12 -0.6925376 0.6072929 -1.1403683 2.543016e-01  
## 13 0.6653603 0.3758795 1.7701424 7.689159e-02  
## 14 0.6360804 3.3964682 0.1872770 8.514669e-01  
## 15 -0.4871436 0.6627649 -0.7350173 4.624354e-01  
## 16 -0.4823724 0.7555671 -0.6384243 5.232878e-01  
## 17 -0.4752182 1.4623948 -0.3249589 7.452542e-01  
## 18 0.4565020 0.2916115 1.5654457 1.176739e-01  
## 19 0.4023233 1.0697394 0.3760947 7.068958e-01  
## 20 -0.3908215 0.3171660 -1.2322301 2.180421e-01  
## var  
## 1 (Intercept)  
## 2 DisciplineStructural  
## 3 code.directorS153  
## 4 code.directorS139  
## 5 code.directorS150  
## 6 Businessmembrane fabricator  
## 7 Businessresources  
## 8 Biz.typepublic  
## 9 code.directorS116  
## 10 Businessinternal  
## 11 mean.peeps  
## 12 Businessartist  
## 13 Job.Type.Primary5. SpecialStructures  
## 14 BusinessNFP  
## 15 Businesscouncil  
## 16 Biz.typeNFP  
## 17 Businessutilities provider  
## 18 ProjEng.PosSenior Professional  
## 19 Biz.sizelocal  
## 20 code.directorS168

BT.lin(type= 'lm', data= all7a[1314:2283,], add.var = 'Job.Type.Primary') %>% slice(1:10)

## Estimate Std. Error t value Pval  
## 1 418.570994 278.9997127 1.5002560 1.339105e-01  
## 2 -3.345428 1.7085815 -1.9580151 5.054714e-02  
## 3 -2.684199 0.6752956 -3.9748497 7.625809e-05  
## 4 2.651889 1.2900128 2.0557079 4.010829e-02  
## 5 -2.629128 1.6655818 -1.5785044 1.148129e-01  
## 6 -2.358034 0.8551360 -2.7574957 5.946503e-03  
## 7 -2.318152 1.4095788 -1.6445705 1.004193e-01  
## 8 -1.620112 2.4051392 -0.6736043 5.007416e-01  
## 9 -1.599509 2.3522003 -0.6800054 4.966817e-01  
## 10 -1.367778 1.3069430 -1.0465478 2.955985e-01  
## var  
## 1 (Intercept)  
## 2 Businessresources  
## 3 DisciplineStructural  
## 4 Biz.typepublic  
## 5 code.directorS139  
## 6 code.directorS150  
## 7 code.directorS153  
## 8 Businessutilities provider  
## 9 code.directorS116  
## 10 Businesslandscape arch

BT.lin(type= 'aov', data= all7a, add.var = 'Job.Type.Primary')

## Df Sum Sq Mean Sq F value Pr(>F)   
## inv.mlsto 1 14 13.54 1.235 0.266584   
## Discipline 3 82 27.20 2.481 0.059395 .   
## client.count 1 5 4.50 0.411 0.521547   
## Business 32 90 2.81 0.256 0.999993   
## Biz.size 4 39 9.85 0.899 0.463833   
## Biz.type 3 65 21.72 1.981 0.114808   
## Year 1 20 20.49 1.870 0.171679   
## Num.days 1 139 138.72 12.657 0.000385 \*\*\*  
## no.users 1 79 79.22 7.229 0.007249 \*\*   
## pc.contracttech 1 6 6.31 0.576 0.448029   
## client.neginv 1 4 4.20 0.383 0.535972   
## client.numinv 1 5 5.34 0.487 0.485430   
## client.totinv 1 1 0.92 0.084 0.771884   
## pc.director 1 10 9.76 0.891 0.345362   
## pc.midtech 1 266 265.78 24.250 9.32e-07 \*\*\*  
## pc.midpro 1 5 4.58 0.418 0.518104   
## pc.gradpro 1 8 7.67 0.700 0.402854   
## pc.seniortech 1 3 3.00 0.274 0.600913   
## pc.seniorpro 1 0 0.01 0.001 0.970914   
## mean.peeps 1 110 109.90 10.027 0.001571 \*\*   
## hours.perday 1 86 85.79 7.828 0.005206 \*\*   
## num.inv 1 2 1.56 0.143 0.705634   
## mean.inv 1 6 6.37 0.581 0.445882   
## num.neginv 1 0 0.02 0.002 0.967685   
## client.meaninv 1 4 4.36 0.398 0.528384   
## code.director 6 236 39.42 3.596 0.001515 \*\*   
## ProjEng.Pos 4 30 7.49 0.683 0.603630   
## Job.Type.Primary 5 40 8.05 0.735 0.597294   
## Residuals 1619 17744 10.96   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 651 observations deleted due to missingness

BT.lin(type= 'aov', data= all7a[1314:2283,], add.var = 'Job.Type.Primary')

## Df Sum Sq Mean Sq F value Pr(>F)   
## inv.mlsto 1 23 22.6 1.187 0.276316   
## Discipline 3 117 38.9 2.045 0.106097   
## client.count 1 3 3.3 0.172 0.678076   
## Business 30 88 2.9 0.154 1.000000   
## Biz.size 4 79 19.7 1.033 0.389102   
## Biz.type 3 137 45.6 2.398 0.066698 .   
## Year 1 11 11.5 0.603 0.437584   
## Num.days 1 170 169.8 8.922 0.002897 \*\*   
## no.users 1 140 139.6 7.335 0.006894 \*\*   
## pc.contracttech 1 12 11.8 0.618 0.432021   
## client.neginv 1 6 6.2 0.326 0.568216   
## client.numinv 1 6 6.4 0.337 0.561939   
## client.totinv 1 26 26.3 1.381 0.240219   
## pc.director 1 42 42.5 2.233 0.135457   
## pc.midtech 1 339 339.0 17.818 2.69e-05 \*\*\*  
## pc.midpro 1 7 6.9 0.364 0.546273   
## pc.gradpro 1 13 12.6 0.661 0.416456   
## pc.seniortech 1 13 12.8 0.673 0.412320   
## pc.seniorpro 1 0 0.0 0.002 0.964585   
## mean.peeps 1 111 111.4 5.857 0.015717 \*   
## hours.perday 1 225 225.5 11.851 0.000604 \*\*\*  
## num.inv 1 3 2.6 0.135 0.713884   
## mean.inv 1 48 47.6 2.502 0.114074   
## num.neginv 1 3 2.9 0.151 0.697267   
## client.meaninv 1 9 9.3 0.490 0.484006   
## code.director 6 218 36.4 1.913 0.076018 .   
## ProjEng.Pos 3 41 13.7 0.720 0.539906   
## Job.Type.Primary 5 33 6.5 0.343 0.886613   
## Residuals 871 16573 19.0   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 23 observations deleted due to missingness

Most influential variables are:

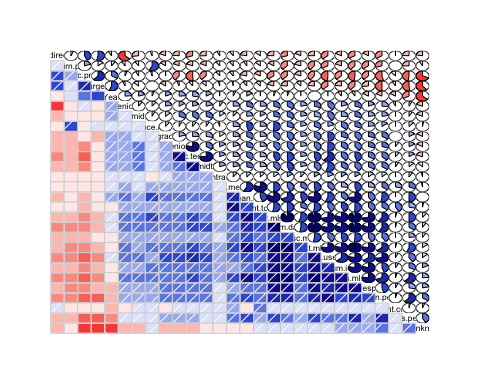
* Business type
* director code
* Discipline
* Biz.type

Excluded variables include:

* Post.Code
* Billing.Type
* Job.Source
* Job.Detail.Primary
* Job.Detail.Secondary
* Job.Type.Primary
* Job.Type.Secondary
* State
* Type
* no.employees
* contact.count
* JD.Primary
* JD.Second
* dist
* timespan (won't know to begin with)
* Num.disc
* Inv.freq
* client.invfreq

# Correlation Analysis - Numeric variables with 'return.pdol' variable

First run Corrgram - corrgram function Run as type 'spearman' - non linear



Now output cor.test results for all numeric variables using both pearson (linear) and kendall (non-linear) to compare. Kendall does not have problems with ties

Output ordered by p value

summ %>% arrange(test, p.val)

## test var corr p.val  
## 1 kendall return.pdol 1.000 0.00e+00  
## 2 kendall pc.pro 0.147 0.00e+00  
## 3 kendall cost.mlsto -0.263 2.24e-81  
## 4 kendall hrs.mlsto -0.260 2.39e-79  
## 5 kendall Num.days -0.257 4.91e-75  
## 6 kendall timespan -0.239 1.17e-66  
## 7 kendall no.users -0.245 1.36e-58  
## 8 kendall mean.peeps -0.213 3.66e-48  
## 9 kendall hours.perday -0.169 2.80e-34  
## 10 kendall dis.sc.mlsto -0.153 4.19e-22  
## 11 kendall Num.disc -0.159 3.80e-21  
## 12 kendall pc.tech -0.147 1.29e-20  
## 13 kendall num.inv -0.136 5.79e-20  
## 14 kendall pc.seniortech -0.139 7.69e-18  
## 15 kendall pc.midtech -0.136 4.02e-17  
## 16 kendall inv.mlsto -0.105 3.14e-14  
## 17 kendall pc.gradpro -0.115 1.32e-12  
## 18 kendall client.numinv -0.097 3.70e-12  
## 19 kendall client.totinv -0.087 2.55e-10  
## 20 kendall num.neginv -0.104 4.70e-10  
## 21 kendall Year 0.069 1.81e-06  
## 22 kendall pc.contracttech -0.080 2.28e-06  
## 23 kendall client.meaninv -0.065 2.64e-06  
## 24 kendall pc.midpro -0.073 7.51e-06  
## 25 kendall pc.seniorpro -0.067 1.17e-05  
## 26 kendall client.neginv -0.063 4.88e-05  
## 27 kendall mean.inv -0.055 6.28e-05  
## 28 kendall pc.unknown -0.062 9.77e-05  
## 29 kendall pc.director 0.048 9.42e-04  
## 30 kendall no.employees 0.028 1.01e-01  
## 31 kendall dist -0.021 1.56e-01  
## 32 kendall Inv.freq -0.017 3.54e-01  
## 33 kendall client.invfreq -0.010 5.18e-01  
## 34 kendall client.count -0.006 6.82e-01  
## 35 pearson return.pdol 1.000 0.00e+00  
## 36 pearson pc.midtech 0.086 3.04e-05  
## 37 pearson hours.perday -0.075 2.60e-04  
## 38 pearson mean.peeps -0.074 3.58e-04  
## 39 pearson no.users -0.073 4.06e-04  
## 40 pearson timespan -0.067 1.14e-03  
## 41 pearson Num.days -0.059 3.94e-03  
## 42 pearson Num.disc -0.051 1.50e-02  
## 43 pearson pc.midpro 0.048 1.97e-02  
## 44 pearson Year 0.036 7.72e-02  
## 45 pearson cost.mlsto -0.036 8.23e-02  
## 46 pearson hrs.mlsto -0.035 8.76e-02  
## 47 pearson pc.tech 0.030 1.43e-01  
## 48 pearson pc.seniortech -0.030 1.51e-01  
## 49 pearson client.neginv -0.027 1.96e-01  
## 50 pearson pc.unknown -0.023 2.76e-01  
## 51 pearson num.neginv -0.022 2.80e-01  
## 52 pearson dis.sc.mlsto -0.022 2.95e-01  
## 53 pearson client.count -0.018 3.88e-01  
## 54 pearson mean.inv -0.017 4.10e-01  
## 55 pearson no.employees 0.017 4.92e-01  
## 56 pearson pc.contracttech -0.014 5.00e-01  
## 57 pearson client.meaninv -0.011 6.05e-01  
## 58 pearson pc.seniorpro -0.010 6.14e-01  
## 59 pearson pc.director -0.010 6.33e-01  
## 60 pearson Inv.freq -0.010 7.05e-01  
## 61 pearson client.invfreq -0.007 7.43e-01  
## 62 pearson inv.mlsto -0.006 7.66e-01  
## 63 pearson pc.gradpro -0.005 8.17e-01  
## 64 pearson client.numinv -0.005 8.26e-01  
## 65 pearson client.totinv -0.004 8.63e-01  
## 66 pearson num.inv -0.002 9.11e-01  
## 67 pearson pc.pro 0.001 9.43e-01  
## 68 pearson dist 0.000 9.95e-01

Output ordered by correlation value - only 'significant' results to 5%

summ %>% filter(p.val<= 0.05) %>% arrange(test, -abs(corr))

## test var corr p.val  
## 1 kendall return.pdol 1.000 0.00e+00  
## 2 kendall cost.mlsto -0.263 2.24e-81  
## 3 kendall hrs.mlsto -0.260 2.39e-79  
## 4 kendall Num.days -0.257 4.91e-75  
## 5 kendall no.users -0.245 1.36e-58  
## 6 kendall timespan -0.239 1.17e-66  
## 7 kendall mean.peeps -0.213 3.66e-48  
## 8 kendall hours.perday -0.169 2.80e-34  
## 9 kendall Num.disc -0.159 3.80e-21  
## 10 kendall dis.sc.mlsto -0.153 4.19e-22  
## 11 kendall pc.pro 0.147 0.00e+00  
## 12 kendall pc.tech -0.147 1.29e-20  
## 13 kendall pc.seniortech -0.139 7.69e-18  
## 14 kendall pc.midtech -0.136 4.02e-17  
## 15 kendall num.inv -0.136 5.79e-20  
## 16 kendall pc.gradpro -0.115 1.32e-12  
## 17 kendall inv.mlsto -0.105 3.14e-14  
## 18 kendall num.neginv -0.104 4.70e-10  
## 19 kendall client.numinv -0.097 3.70e-12  
## 20 kendall client.totinv -0.087 2.55e-10  
## 21 kendall pc.contracttech -0.080 2.28e-06  
## 22 kendall pc.midpro -0.073 7.51e-06  
## 23 kendall Year 0.069 1.81e-06  
## 24 kendall pc.seniorpro -0.067 1.17e-05  
## 25 kendall client.meaninv -0.065 2.64e-06  
## 26 kendall client.neginv -0.063 4.88e-05  
## 27 kendall pc.unknown -0.062 9.77e-05  
## 28 kendall mean.inv -0.055 6.28e-05  
## 29 kendall pc.director 0.048 9.42e-04  
## 30 pearson return.pdol 1.000 0.00e+00  
## 31 pearson pc.midtech 0.086 3.04e-05  
## 32 pearson hours.perday -0.075 2.60e-04  
## 33 pearson mean.peeps -0.074 3.58e-04  
## 34 pearson no.users -0.073 4.06e-04  
## 35 pearson timespan -0.067 1.14e-03  
## 36 pearson Num.days -0.059 3.94e-03  
## 37 pearson Num.disc -0.051 1.50e-02  
## 38 pearson pc.midpro 0.048 1.97e-02

From this list it can be concluded what the most influential variables are. For the purposes of return.pdol prediction, cost.mlsto, hrs.mlsto, Num.days, dis.sc.mlsto, num.neginv, Year (won't help in the future) cannot be known.

This leaves the following important variables:

|  |  |  |
| --- | --- | --- |
| var | p.val | corr |
| no.users | 0.00e+00 | -0.245 |
| timespan | 0.00e+00 | -0.239 |
| mean.peeps | 0.00e+00 | -0.213 |
| hours.perday | 0.00e+00 | -0.169 |
| Num.disc | 0.00e+00 | -0.159 |
| pc.pro | 0.00e+00 | 0.147 |
| pc.tech | 0.00e+00 | -0.147 |
| pc.seniortech | 0.00e+00 | -0.139 |
| pc.midtech | 0.00e+00 | -0.136 |
| num.inv | 0.00e+00 | -0.136 |
| pc.gradpro | 0.00e+00 | -0.115 |
| inv.mlsto | 0.00e+00 | -0.105 |
| client.numinv | 0.00e+00 | -0.097 |
| client.totinv | 0.00e+00 | -0.087 |
| pc.contracttech | 2.30e-06 | -0.080 |
| pc.midpro | 7.50e-06 | -0.073 |
| pc.seniorpro | 1.17e-05 | -0.067 |
| client.meaninv | 2.60e-06 | -0.065 |
| client.neginv | 4.88e-05 | -0.063 |
| pc.unknown | 9.77e-05 | -0.062 |
| mean.inv | 6.28e-05 | -0.055 |
| pc.director | 9.42e-04 | 0.048 |

# Strength of association with Categorical variables