Factor model of cross-sectional US stock returns

(linear regression diagnostics)

Problem statement

Main **metric of interest**:

(rate of) **return** = 100* (yesterday close price – today price) / today price

Stock universe – Russell 3000 Index.
US Stocks only. Representative of all US liquid investable stocks
Total US stocks ~ 13,000
Total stocks in the world ~ 70,000

Dates: main analysis done for March 24, 2017. Data available for last year.

Main goal:
Explain the returns of Russell 3000 stocks on a given date as a sum of factor returns on a given date

Cross-sectional factor models

Stock return_i ~ sum(factor_loading * factor_return)

Two types of factors:

Style factors:

- **Size (large vs small)**
- **Value vs growth**
- Momentum

Industry factors: GICS (Global Industry Classification Standard)

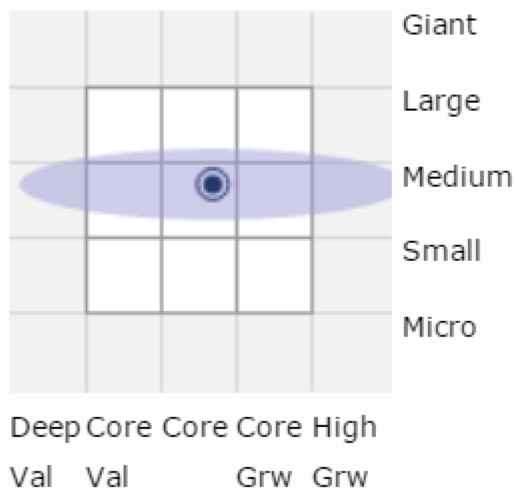
Factors: style

Vanguard Mid Cap Index Institutional VMCIX

Category
Mid-Cap Blend

Investment Style
Mid Blend

Style Map VMCIX



https://awrd.morningstar.com/SBT/Tools/MR/Default.aspx?fullversion=1&ticker=VMCIX

GICS

Classification[4] [edit]

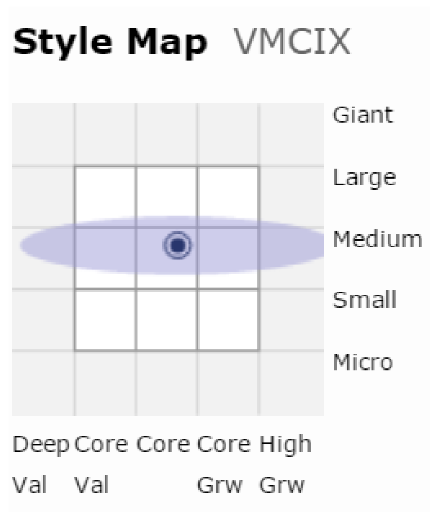
- 11 Sectors
- 24 Industry Groups
- 68 Industries
- 157 Sub-Industries

Sector		Industry Group		Industry		Sub-Industry	
10	Energy	1010	Energy	101010	Energy Equipment & Services	10101010	Oil & Gas Drilling
						10101020	Oil & Gas Equipment & Services
				101020	Oil, Gas & Consumable Fuels	10102010	Integrated Oil & Gas
						10102020	Oil & Gas Exploration & Production
						10102030	Oil & Gas Refining & Marketing
						10102040	Oil & Gas Storage & Transportation
						10102050	Coal & Consumable Fuels

DataFrame used in regression

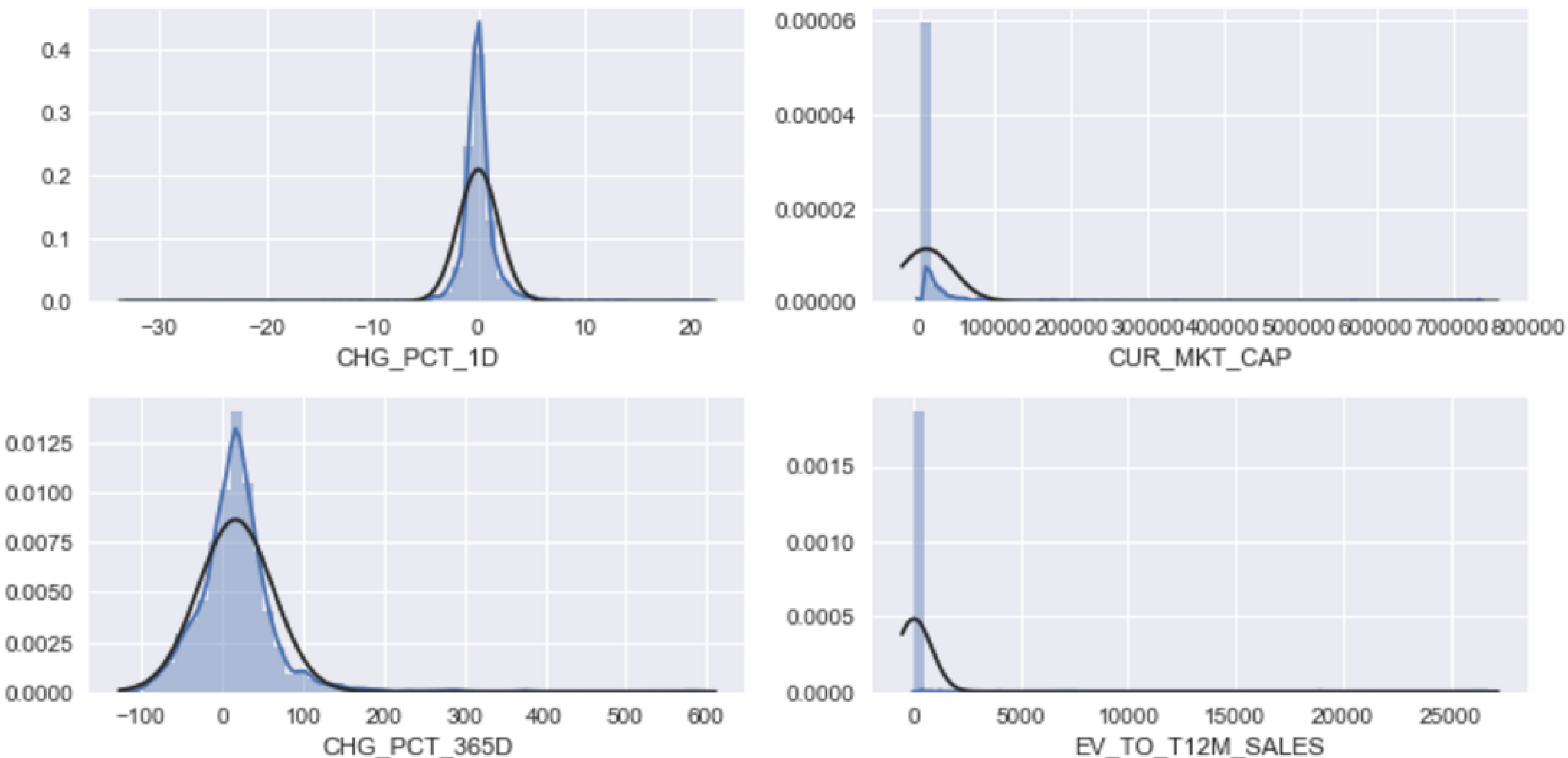
	CHG_PCT_1D	CHG_PCT_365D	CUR_MKT_CAP	EV_TO_T12M_SALES	GICS_1	GICS_2	GICS_3	GICS_4
A	-0.19	50.39	17107.73	3.66	35	3520	352030	35203010
AA	-2.25	46.90	5997.71	0.51	15	1510	151040	15104010
AAC	2.18	-70.18	189.05	2.18	35	3510	351020	35102020

smf.ols(formula=CHG_PCT_1D~CHG_PCT_365D + LOG_CUR_MKT_CAP + LOG_SALES_TO_EV +C(GICS_1)



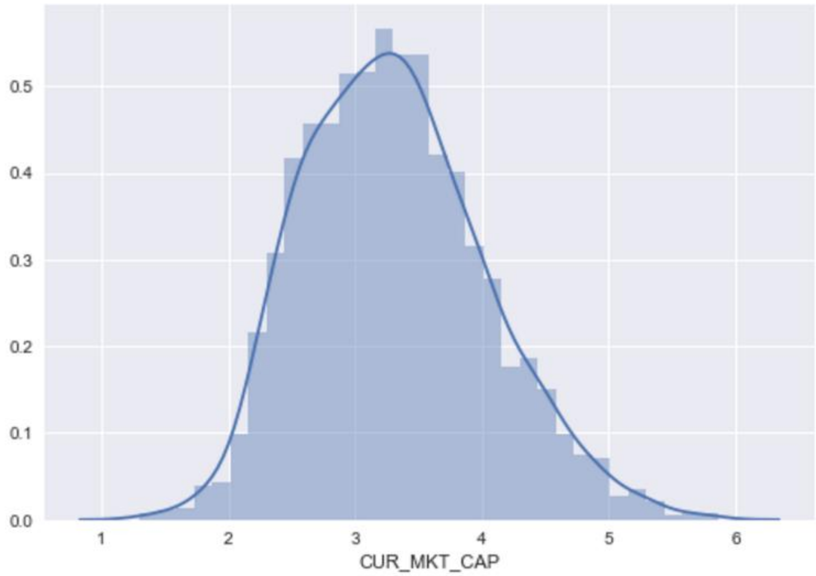
Ford EV/Sales=0.24 IBM EV/Sales=2.4 Snapchat EV/Sales = 25

Variable transformations

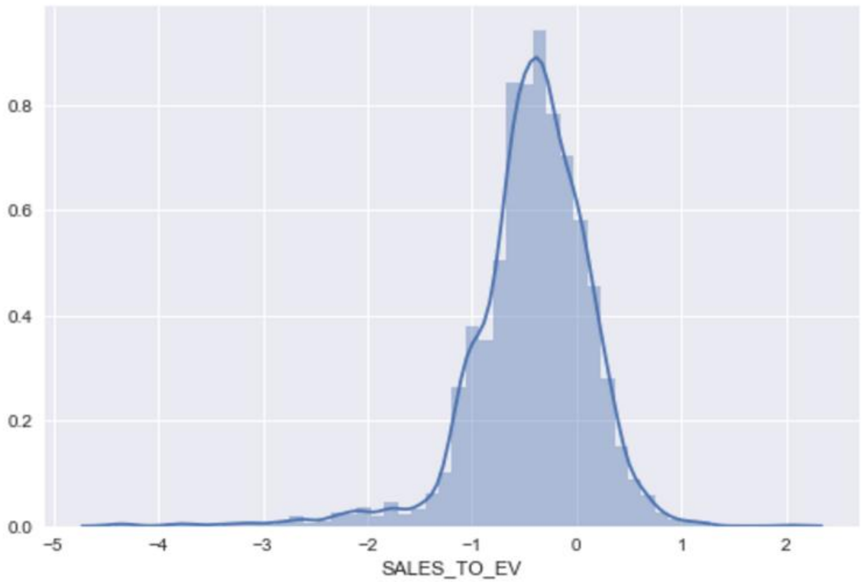


Variable transformations

```
sns.distplot(np.log10(df['CUR_MKT_CAP']));
```




```
sns.distplot(np.log10(df['SALES_TO_EV']));
```



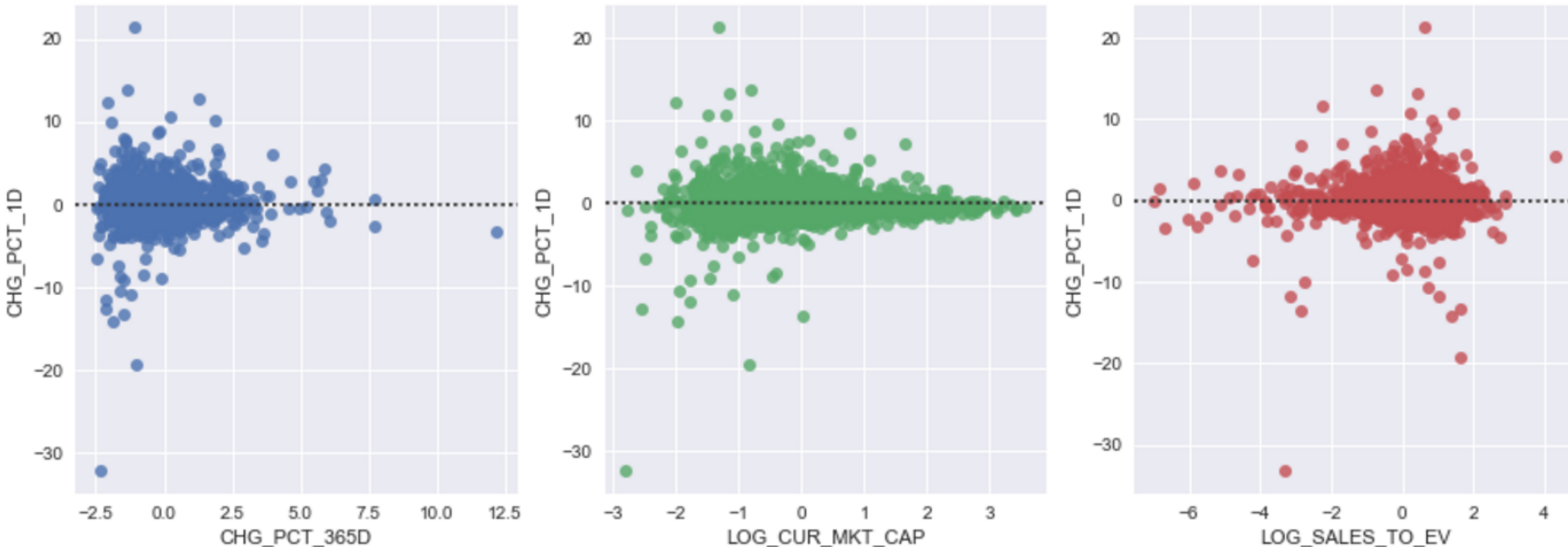
Three-factor model (no industry factors)

smf.ols(formula='CHG_PCT_1D ~ CHG_PCT_365D + LOG_CUR_MKT_CAP + LOG_SALES_TO_EV'

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             CHG_PCT_1D   R-squared:                       0.020
Model:                            OLS   Adj. R-squared:                  0.019
Method:                 Least Squares   F-statistic:                     17.35
Date:                Wed, 29 Mar 2017   Prob (F-statistic):           3.78e-11
Time:                        21:59:57   Log-Likelihood:                -5235.5
No. Observations:                2554   AIC:                         1.048e+04
Df Residuals:                    2550   BIC:                         1.050e+04
Df Model:                           3
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept           0.0032      0.037      0.085      0.932      -0.070       0.076
CHG_PCT_365D        0.1970      0.038      5.157      0.000       0.122       0.272
LOG_CUR_MKT_CAP     0.0296      0.038      0.775      0.438      -0.045       0.105
LOG_SALES_TO_EV    -0.1647      0.037     -4.417      0.000      -0.238      -0.092
==============================================================================
Omnibus:                     1497.381   Durbin-Watson:                   1.993
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           326945.264
Skew:                          -1.650   Prob(JB):                         0.00
Kurtosis:                      58.330   Cond. No.                         1.26
==============================================================================
```

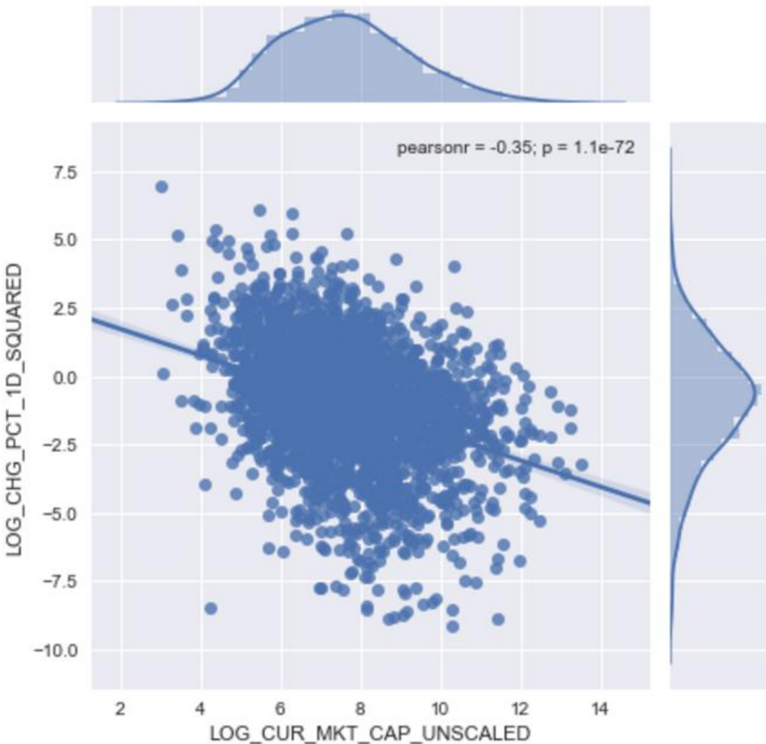
Residuals: conditional heteroscedasticity

```
sns.residplot(y='CHG_PCT_1D', x=col, data=df_norm, ax=axes[i])
```

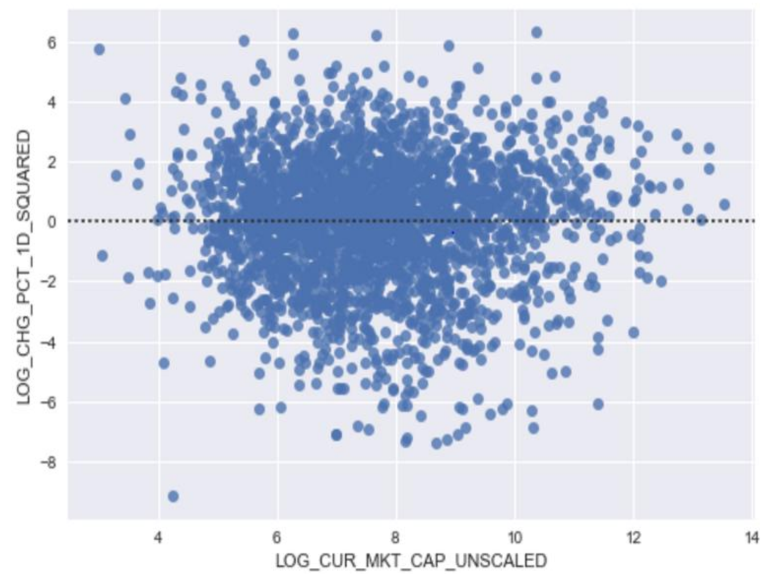


Squared residual ~ market cap

LOG (CHG_PCT_1D^2) ~ -0.48 * LOG_CUR_MKT_CAP




```
sns.residplot(y='LOG_CHG_PCT_1D_SQUARED', x='LOG_CUR_MKT_CAP_UNSCALED',
```



Weighted least squares (WLS)

```
weights = pow( df_norm['CUR_MKT_CAP'].values, 0.48)
sm.WLS(Y,X, weights=weights )
```

Dep. Variable:	y	**R-squared:**	0.025
Model:	WLS	**Adj. R-squared:**	0.024
Method:	Least Squares	**F-statistic:**	21.95
Date:	Wed, 29 Mar 2017	**Prob (F-statistic):**	4.96e-14
Time:	22:00:12	**Log-Likelihood:**	-4684.3
No. Observations:	2554	**AIC:**	9377.
Df Residuals:	2550	**BIC:**	9400.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.0399	0.033	1.227	0.220	-0.024	0.104
x1	0.0875	0.029	3.020	0.003	0.031	0.144
x2	-0.0597	0.023	-2.605	0.009	-0.105	-0.015
x3	-0.2200	0.031	-7.101	0.000	-0.281	-0.159

Omnibus:	764.916	**Durbin-Watson:**	2.000
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	28818.487
Skew:	0.717	**Prob(JB):**	0.00
Kurtosis:	19.394	**Cond. No.**	2.33

Adding GICS categories to the WLS model

smf.wls('CHG_PCT_1D_WIN~CHG_PCT_365D+LOG_CUR_MKT_CAP+LOG_SALES_TO_EV +
+ **C(GICS_1)**'

Dep. Variable:	CHG_PCT_1D_WIN	**R-squared:**	0.129
Model:	WLS	**Adj. R-squared:**	0.125
Method:	Least Squares	**F-statistic:**	29.00
Date:	Wed, 29 Mar 2017	**Prob (F-statistic):**	2.57e-67
Time:	22:26:01	**Log-Likelihood:**	-4291.6
No. Observations:	2554	**AIC:**	8611.
Df Residuals:	2540	**BIC:**	8693.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-0.1242	0.092	-1.343	0.179	-0.305	0.057
C(GICS_1)[T.15.0]	-0.8381	0.134	-6.273	0.000	-1.100	-0.576
C(GICS_1)[T.20.0]	-0.1499	0.110	-1.367	0.172	-0.365	0.065
C(GICS_1)[T.25.0]	0.3633	0.107	3.403	0.001	0.154	0.573
C(GICS_1)[T.30.0]	0.1283	0.125	1.030	0.303	-0.116	0.373
C(GICS_1)[T.35.0]	0.7405	0.108	6.859	0.000	0.529	0.952
C(GICS_1)[T.40.0]	-0.0766	0.115	-0.665	0.506	-0.302	0.149
C(GICS_1)[T.45.0]	0.3483	0.105	3.304	0.001	0.142	0.555
C(GICS_1)[T.50.0]	0.6529	0.204	3.195	0.001	0.252	1.054
C(GICS_1)[T.55.0]	0.5415	0.142	3.809	0.000	0.263	0.820
C(GICS_1)[T.60.0]	-0.0151	0.124	-0.122	0.903	-0.259	0.228
CHG_PCT_365D	0.1085	0.026	4.215	0.000	0.058	0.159
LOG_CUR_MKT_CAP	-0.0813	0.020	-4.044	0.000	-0.121	-0.042
LOG_SALES_TO_EV	-0.1576	0.030	-5.280	0.000	-0.216	-0.099

C(GICS_1) **R^2** 0.13

C(GICS_2) **R^2** 0.15

C(GICS_3) **R^2** 0.20

C(GICS_4) **R^2** 0.26

Deeper GICS levels

C(GICS_1) **R^2** 0.13

C(GICS_2) **R^2** 0.15

C(GICS_3) **R^2** 0.20

C(GICS_4) **R^2** 0.26

Classification[4]

- 11 Sectors
- 24 Industry Groups
- 68 Industries
- 157 Sub-Industries

Five-fold cross-validation to determine the best level of GICS

GICS_1, mse=3.44754086155, std=1.1499772895
GICS_2, mse=3.42538655171, std=1.1402301803
CRASH!!!

Future directions

- Cross-validate, select the right GICS variables
- Add more style factors (Bloomberg model uses 10)
- Add events: earnings announcement, etc.
- Try non-linear models: Random Forest Regressor

Conclusions

- It is important to transform the variables correctly.
- Regressions diagnostics is extremely important
- Linear regression model requires a lot of work to be specified correctly as adding or removing regressors changes the significance of the original regressors