Received: 7 February 2020 | Revised: 28 June 2020 | Accepted: 9 July 2020

DOI: 10.1111/test.12233

ORIGINAL ARTICLE

WILEY

Multiple linear regression: Identify potential health care stocks for investments using out-of-sample predictions

Thinh Kieu[1] | **Phong Luu**[1] | **Noah Yoon**[2]

[1]Department of Mathematics, University of North Georgia, Oakwood, Georgia

[2]Columbus High School, Columbus, Georgia

Correspondence
Phong Luu, Department of Mathematics, University of North Georgia, Gainesville Campus, 3820 Mundy Mill Rd, Oakwood, GA 30566.
Email: phong.luu@ung.edu

Abstract

College-level statistics courses emphasize the use of the coefficient of determination, R-squared, in evaluating a linear regression model: higher R-squared is better. This often gives students an impression that higher R-squared implies better predictability since textbooks tend to use sample data to support the theory and students rarely have an opportunity to work on real data. In this paper, health care stocks are used as predictors and the result demonstrates that high R-squared does not necessarily mean high predictability and that multiple linear regression can be used in the study of data behavior. In particular, by learning the pattern of the near and far out-of-sample-prediction errors for different time periods throughout a dataset, the near out-of-sample prediction errors can be used to control the prediction errors and identify a subset of predictors that can well reflect the trend of S&P 500.

KEYWORDS

health care stocks, multiple linear regression, out-of-sample prediction, R-squared, S&P 500, teaching statistics

1 | INTRODUCTION

Standard & Poor's 500 Index (S&P 500) is a market-capitalization-weighted index of the 500 largest U.S. publicly traded companies. The index is widely regarded as the best gauge of large-cap U.S. equities [4]. In this paper, S&P 500 will be used as a measure of the U.S. economy. According to the 2018 annual results of the National Association of Insurance Commissioners [6], the health insurance industry continued its tremendous growth trend as it experienced a significant increase in net earnings to $23.4 billion and an increase in the profit margin to 3.3% in 2018 compared to net earnings of $16.1 billion and a profit margin of 2.4% in 2017. For traders who are interested in investing in the health care industry, it is critical for them to identify health care companies which can reflect the economic health of the U.S. S&P 500, since there is a good chance that these are well-managed and high quality ones. Subsequent analyses of these stocks are necessary to see whether they offer appropriate potential returns.

Using the history of daily adjusted prices of the health care industry, the project sets out to use multiple linear regression to produce a subset of health care stocks which can be used as an indicator of the market trend (S&P 500) for a short term. While S&P 500 tracks large-cap companies, small-cap ones can collectively affect its trend. Moreover, smaller companies have more room to grow, giving shareholders the opportunity to realize substantial gains on these investments [5]. Proponents of the efficient market hypothesis conclude that, because of the randomness of the market, investors could do better by investing in a low-cost and passive portfolio [3]. Leaning too heavily on large-cap stocks can also put traders in a precarious position. "Most people consider large caps to be safer, which is often true," Yoder said. "But in the last two recessions, small caps actually held up better than large caps." In other words, people with a portfolio full of

mostly large-cap stocks lost more money than those with some small caps mixed in, even though it is the riskier choice. In the end, it is crucial to have a mix of both small caps and large caps in one's portfolio [5].

We will identify a subset of potential stocks for investments, among the total of 250 health care stocks, using multiple linear regression. The resulting model has high predictability and satisfies all assumptions of the multiple linear regression method.

The health care industry is large and using too many predictors in multiple linear regression can overload the prediction system. For this reason, one of the initial and most challenging steps is determining a manageable number of predictors, which has the strongest predicting ability for S&P 500. In Enke et al. [1], the multiple linear regression analysis is performed on 25 finance and economic variables to both reduce the dimensionality of the variable set and identify which variables have a strong relationship with the market price of the S&P 500 Index for the subsequent testing period. Other attempts at predicting S&P 500 using multiple linear regression include Seethalakshmi [7] and Smith et al. [8]. All of these papers give models with a high coefficient of determination, R-squared, and use the high R-squared as an indication of a good model. This paper shows that high R-squared is not very useful for the considered dataset; hence, it is important to study the data behavior when using multiple linear regression to build a high predictability model.

2 | DATA DESCRIPTION

The data contains daily adjusted prices of S&P 500 and 250 health care companies from October 1, 2018 through October 7, 2019 (256 observations), downloaded from YAHOO FINANCE. The stock return of each stock will be computed by taking the adjusted closing stock price on day $i + 1$ and subtracting the adjusted closing price of the previous day. This computation is shown in the equation below.

$$\mathrm{SR}_i = P_{i+1} - P_i$$

Note that the stock returns contain 255 observations (Table 1).

3 | METHODOLOGY

We use the first 245 observations for building models, the next five observations for selecting a potentially best model, and the last five observations for testing the selected model. For the given time period October 1, 2018 to September 30, 2019 or the first 250 observations of the stock returns, we wish to identify subsets of 250 health care companies that can predict accurately the trend of SP500. The general idea is to generate random subsets of size 10 from the 250 health care companies and run multiple linear regression using SP500 as the response variable and each subset as predictors.

The statistical model has the following form.

$$\mathrm{SP500} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{10} X_{10} + \varepsilon,$$

where X_1, X_2, ..., X_{10} are elements of $\{H_1, H_2, \ldots, H_{250}\}$, and ε is a random error that is normally distributed with mean zero and constant variance.

TABLE 1 Description of variables

Variable	Dependence	Type	Description
SP500	Response	Continuous	Stock return for S&P 500.
$H_1, H_2, \ldots, H_{250}$	Predictors	Continuous	Stock returns for 250 health care companies[a]

[a]250 health care company symbols include JNJ, UNH, MRK, PFE, NVS, MDT, ABT, AMGN, NVO, ABBV, AZN, TMO, SNY, LLY, GSK, DHR, CVS, BMY, GILD, CELG, SYK, CI, ANTM, BDX, ISRG, TAK, AGN, BSX, ZTS, VRTX, WBA, BIIB, EW, HCA, ILMN, HUM, BAX, PHG, REGN, ZBH, IQV, MCK, CNC, ALXN, A, VEEV, FMS, IDXX, CERN, ALGN, RMD, SGEN, DXCM, GRFS, SNN, INCY, ABC, MTD, LH, CAH, TFX, WCG, COO, GMAB, DGX, WAT, BMRN, HOLX, STE, XRAY, UHS, GLPG, VAR, ALNY, TEVA, PODD, WST, BIO, NBIX, HSIC, PKI, BHC, QGEN, DVA, NVCR, MYL, IONS, MOH, ICLR, JAZZ, MASI, ABMD, TECH, SRPT, BRKR, CTLT, ACAD, EHC, CHE, CRL, HRC, RDY, PRGO, PRAH, CGC, HAE, HZNP, TDOC, PEN, SYNH, GMED, MDCO, ASND, AMED, IART, EXEL, ARWR, CBPO, RGEN, HQY, ONCE, LHCG, UTHR, BLUE, LIVN, ICUI, WMGI, BPMC, NKTR, NUVA, TARO, MRTX, NEOG, GBT, FGEN, NVRO, OMCL, ALKS, THC, CMD, CNMD, NTRA, ICPT, SEM, ACHC, INOV, HALO, AMN, QDEL, HMSY, PTCT, EBS, AGIO, ITGR, PINC, ENSG, GKOS, QURE, MD, RARE, PTLA, CBM, NVTA, INSM, MGLN, PBH, MYGN, IRWD, PCRX, PDCO, HCSG, MDRX, AIMT, AVNS, INGN, NRC, CSII, MNTA, MMSI, BEAT, EPZM, USPH, RGNX, ADUS, ATRI, BKD, INVA, ENTA, NXGN, SGMO, NTUS, TVTY, KPTI, XON, SPPI, HNGR, LMNX, CRY, HSTM, CDXS, EVH, ACHN, AMPH, OFIX, ANIK, PACB, PRSC, ATRA, DBVT, COLL, SGRY, MESO, ANGO, CYTK, BDSI, ITCI, SRDX, CUTR, XBIT, RAD, OSUR, CLVS, DERM, PETS, CCXI, PGNX, ATEC, GTS, PRTA, MGNX, SIEN, OMI, CPSI, BLU, CBAY, UTMD, VIVO, ASMB, AMAG, PBYI, LCI, CBMG, IVC, ARA, CTMX, SPNE, KMDA, CHMA, AXGT, ARAV.

TABLE 2 1672 subsets of size 10 of health care companies with R-squared of at least 0.7 and the corresponding near out-of-sample prediction error E_1 and far out-of-sample prediction error E_2

	Predictors	E_1	E_2	R-Squared
1	MOH + WBA + BAX + CELG+SGMO+TECH+WMGI+AIMT+CRL + TMO	0.5665	0.1908	0.7755
2	RAD+CERN+EBS + MRK + MESO+ZTS + MDT + PBYI+PKI + SRPT	1.1063	0.4552	0.775
3	PKI + AMPH+ITGR+ZTS + IQV + HSIC+ANGO+BLU + ANIK+CAH	0.4913	0.2505	0.7745
4	PODD+MESO+CHE + TMO + ILMN+ABT + BMRN+PRGO+ALKS+IART	0.5147	0.2426	0.7718
5	NVTA+CERN+EBS + CBMG+IDXX+TVTY+JNJ + TMO + BIO + SGRY	0.6631	0.3714	0.7698
6	ACHC+SGMO+ZTS + WBA + MRTX+IDXX+STE + ABMD+NVRO+EW	0.2306	0.3011	0.7698
7	KPTI+LH + SIEN+MASI+IDXX+PACB+WBA + ZTS + NTUS+ANTM	0.43	0.2708	0.7676
8	HQY + ZTS + PRTA+TMO + INGN+ALKS+GMAB+SNY + CVS + NVRO	0.9139	0.2521	0.7661
9	UHS + BMRN+IDXX+MCK + ABT + ILMN+ICLR+ABMD+TMO + TVTY	0.3409	0.2056	0.7651
10	CGC + DGX + AMGN+WBA + TMO + NEOG+HALO+TVTY+HAE + BMRN	0.9185	0.1566	0.7651
11	MRK + CPSI+ALXN+WAT + IDXX+MDRX+THC + MYL + HSIC+GBT	0.5786	0.333	0.765
12	ANTM+WBA + BKD + CYTK+NUVA+AGN + SGMO+ISRG+ZTS + ARA	0.4424	0.3015	0.7649
13	AMAG+DHR + CRL + PRSC+LIVN+NBIX+QGEN+TEVA+WBA + HRC	0.6035	0.2038	0.7647
14	IDXX+CHE + AGIO+CMD + TMO + ILMN+NXGN+BHC + LLY + MYL	0.4282	0.2966	0.7641
15	MDCO+CHE + HRC + CRL + DGX + GLPG+QDEL+TEVA+ILMN+WBA	0.4415	0.5136	0.7638
16	MCK + VAR + ABT + PKI + IQV + MD + KPTI+ENSG+CBM + BAX	0.4591	0.3526	0.7637
17	BDX + MRK + RDY + CVS + ACHN+CHE + TEVA+QGEN+PEN+PGNX	0.6954	0.369	0.7635
18	PRAH+JNJ + NEOG+PHG + VIVO+CYTK+VAR + PKI + EVH + TEVA	0.7626	0.3611	0.7633
19	ZTS + INCY+NBIX+EBS + GTS + IDXX+WBA + RAD+LMNX+CERN	0.5133	0.3497	0.7632
20	PRSC+PKI + ABC + ZTS + NVTA+HCA + AGN + HSIC+ARWR+REGN	0.542	0.3009	0.7621
21	DHR + ANGO+INSM+ABT + VEEV+GKOS+WBA + MGNX+EBS + LIVN	0.6246	0.3982	0.7617
22	ABT + BRKR+VAR + HRC + PRAH+WBA + ACAD+LH + PRTA+WMGI	0.3669	0.2436	0.7613
23	NTRA+TECH+ISRG+CELG+MYL + PKI + SYNH+GMED+AMPH+CHE	0.39	0.5048	0.7606
24	INVA+CRL + HAE + TFX + TEVA+ICPT+HOLX+CHE + LH + PKI	0.4382	0.5592	0.76
25	IDXX+ABC + ZTS + AMAG+MGLN+BSX + WBA + NEOG+OMI + BLUE	0.4834	0.2915	0.7597
26	MNTA+BEAT+MDRX+BDX + UHS + PKI + BMY + AMN + EPZM+TDOC	0.5081	0.2535	0.759
27	INVA+AGN + CNC + CERN+PRAH+PKI + ISRG+SGRY+BMY + IONS	0.6382	0.2984	0.759
28	PRGO+ZTS + PKI + MYL + HCSG+PRAH+MD + INVA+NVTA+IONS	0.7163	0.2107	0.7585
29	CLVS+CAH + ABBV+SIEN+CHE + GTS + ALKS+CDXS+UNH + TMO	0.8917	0.2129	0.7584
30	CBMG+TMO + CHE + PRTA+CGC + ALGN+SYK + ANGO+HNGR+MD	0.5004	0.3711	0.7583
31	EBS + EW + ALXN+MRK + TECH+WBA + AMED+NXGN+SYNH+CERN	0.2497	0.2319	0.7583
32	PKI + TFX + A + ZBH + MDT + CCXI+CTLT+EPZM+RARE+ZTS	0.8116	0.3372	0.7565
33	VAR + ISRG+IDXX+FMS + JNJ + PETS+VIVO+ZTS + PODD+ALGN	0.6774	0.3817	0.7563
34	GLPG+RARE+CYTK+RAD+IDXX+BHC + ABBV+ZTS + NKTR+PETS	0.8968	0.3056	0.7557
35	WBA + KPTI+BPMC+MTD + QGEN+DGX + PRAH+QDEL+DERM+SRPT	0.2351	0.2252	0.7557
36	ALNY+ZTS + AMPH+BEAT+NTUS+PKI + QDEL+BHC + NUVA+CI	0.7737	0.2005	0.7555
37	NTRA+NVS + CHMA+ILMN+ABT + WBA + CAH + CERN+LHCG+IQV	0.6683	0.2226	0.7554
38	VEEV+WBA + PFE + CPSI+PEN+MTD + CHMA+ISRG+QDEL+SRDX	0.3783	0.2637	0.7551
39	PHG + TDOC+CI + CVS + CRY+ABT + HAE + ONCE+BMY + BAX	0.9872	0.2277	0.7548
40	AMN + ZTS + ISRG+QGEN+PRGO+BLU + ABC + UHS + ADUS+A	0.7763	0.185	0.7543
1668	LCI + ALGN+ABMD+SPPI+WST + CTMX+CAH + CRL + BAX + GKOS	0.4375	0.2311	0.7001
1669	BEAT+HSIC+DXCM+TFX + TMO + ALGN+GILD+INCY+PODD+MOH	0.8239	0.2564	0.7

(Continues)

TABLE 2 (Continued)

	Predictors	E_1	E_2	R-Squared
1670	UNH + RARE+ICLR+ITCI+CHE + GTS + OSUR+NEOG+SGMO+MDT	1.219	0.5256	0.7
1671	STE + THC + AMGN+CYTK+QDEL+TEVA+IQV + ICLR+ANIK+WST	1.3097	0.531	0.7
1672	HQY + A + TMO + CUTR+MCK + COLL+CHMA+NTRA+BEAT+AVNS	0.5452	0.1387	0.7

There are two steps in building a model with a predetermined error threshold .

Step 1. Perform multiple linear regression on SP500 against a random subset of 10 predictors using the observations from 1 through 245 to obtain a model.

Step 2. Choose the model in Step 1 which gives a small potential prediction error δ.

Since the observations from 246 through 250 are known, we can use the model in Step 1 to calculate the percentage prediction errors for this time period (246-250). The mean of the absolute values of these percentage errors is denoted by E_1, which is called the near out-of-sample prediction error. For $i = 246 \ldots 250$,

$$PE_i = \frac{\text{SP500}_{\text{PREDICTED}_i} - \text{SP500}_{\text{OBSERVED}_i}}{\text{SP500}_{\text{OBSERVED}_i}}$$

$$E_1 = \frac{|\text{PE}_{246}| + |\text{PE}_{247}| + |\text{PE}_{248}| + |\text{PE}_{249}| + |\text{PE}_{250}|}{5}$$

Select the set of predictors which gives $E_1 < \delta$.

3.1 | R-squared vs predictability

In this section, the confidence intervals of the prediction errors are calculated for multiple linear regression models which have high R-squared values. We generate 10 000 random subsets of predictors of size 10, and run multiple linear regression on SP500 against each subset using the first 245 observations of the data. Among these, 1672 models give R-squared values of at least 0.7. We also calculate the corresponding average absolute percentage errors of the near out-of-sample predictions (using observations from 246 through 250) E_1's and those of the far out-of-sample predictions (using observations from 251 through 255) called E_2's (Table 2).

$$99\%\text{confidence interval of } E_1 = (0.6596, 0.6882).$$

$$99\%\text{confidence interval of } E_2 = (0.3380, 0.3524).$$

For R-squared of at least 0.7, we are 99% confident that the near out-of-sample prediction error is between 65.96% and 68.82%, and the far out-of-sample prediction error is between 33.8% and 35.24%. Thus, this error rate is too high.

3.2 | Learn the data behavior

We will use multiple linear regression to examine the near out-of-sample prediction errors and the far out-of-sample prediction errors for different time periods throughout the dataset. In particular, we will calculate the probability of $E_2 < E_1$ for a random time period. If the probability is high, it will be reasonable to use E_1 to control E_2. To estimate this probability, the time periods of the first 120, 125, 130, 135, 140, 145, ..., 235, and 240 observations will be used.

For each time period, the following are performed. The time period of the first 120 observations will be used for demonstration of the procedure. Other time periods are quite similar.

1. One thousand random groups of 10 predictors are generated from the 250 health care companies. These groups will be used in the next step.
2. A multiple linear regression is run on SP500 against each group of 10 predictors using observations from 1 through 120, and the corresponding E_1 (using observations from 121 through 125) and E_2 (using observations from 126 through 130) are calculated. Let

$$D = E_2 - E_1.$$

3. Calculate the 99% confidence interval for D. Note D contains 1000 values

The results are summarized in the following table.

It has been shown in Table 3 that 15 out of 25 time periods give negative confidence intervals, so we are 99% confident the approximation of the probability of $E_2 < E_1$ is 0.6. Hence, it is fairly reasonable to use E_1 to control the prediction error E_2. It has also been noted from Table 3 some confidence intervals are very high in absolute values. This stemmed from the fact that we did not

TABLE 3 Time periods and the corresponding confidence intervals of the differences $E_2 - E_1$

Time period	Lower bound	Upper bound
1-120	−42.2132	−34.9792
1-125	40.1265	43.3135
1-130	−39.935	−36.8397
1-135	0.9973	1.3454
1-140	−2.5347	−2.1959
1-145	−1.1208	−1.0195
1-150	−0.1147	−0.0552
1-155	0.3456	0.4441
1-160	−0.3076	−0.1806
1-165	1.6581	1.9784
1-170	−0.924	−0.5916
1-175	−0.0962	0.0447
1-180	−1.1977	−1.0741
1-185	0.4047	0.5008
1-190	2.0676	2.5127
1-195	−3.2443	−2.7284
1-200	0.6676	0.7622
1-205	−0.013	0.2773
1-210	−1.0398	−0.7828
1-215	1.2046	1.3586
1-220	−0.8185	−0.6153
1-225	16.7526	18.9393
1-230	−15.184	−12.8421
1-235	25.1256	28.7756
1-240	−33.0135	−29.304

remove random subsets of predictors, which produced high values of E_1, from the samples used to calculate the confidence intervals. In the next section, we will find potential predictors by restricting the values of E_1.

3.3 | Identify a "best" subset of predictors

In this section, potential models are identified by setting E_1 small. In particular, by setting $E_1 < 0.2$, multiple linear regression can find 50 random subsets of size 10 of health care companies. The corresponding prediction errors E_2's are also calculated for evaluation of predictability (Table 4).

99% confidence interval for E_2 is (0.1512, 0.2013). Hence, by setting $E_1 < 0.2$, we are 99% confident that the prediction error is between 15.12% and 20.13%.

Now by using multiple linear regression and setting $E_1 < 0.1$, we obtain the 10 predictors: *SYNH*, *MRK*, *PODD*, *VIVO*, *MGLN*, *AVNS*, *PDCO*, *WST*, *MTD*, and *COO*, which give a model with $E_1 = 0.0886$ and $E_2 = 0.0855$. For convenience, we call these predictors X_1, X_2, ..., X_{10}. The corresponding model is called the full model.

3.4 | Simplify the full model

First, we will test if the regression on SP500 against X_1, X_2, ..., X_{10} explains a significant proportion of the variability in SP500. We will conduct the overall F test. Formally, we state the null and alternative hypotheses as

- $\mathbf{H_0}$: The regression on X_1, X_2, ..., X_{10} does not explain a significant proportion of the variability in SP500.
- $\mathbf{H_A}$: The regression on X_1, X_2, ..., X_{10} explain a significant proportion of the variability in SP500.

Run the following R codes:

```
full = "SP500 ~ SYNH + MRK + PODD + VIVO
    + MGLN + AVNS + PDCO + WST + MTD + COO"

full.lm = lm(formula = full, data = df[1 : 245,])

summary(full.lm)
```

The simplified output is as follows:
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−0.1072	1.0830	−0.099	0.921221
SYNH	3.4874	0.9447	3.691	0.000278 ***
MRK	5.6580	1.3828	4.092	5.90e-05 ***
PODD	0.1651	0.4087	0.404	0.686594
VIVO	0.3138	2.8704	0.109	0.913038
MGLN	1.7413	0.7864	2.214	0.027763 *
AVNS	2.2920	1.0489	2.185	0.029865 *
PDCO	11.8432	2.7396	4.323	2.28e-05 ***
WST	2.7599	0.7023	3.930	0.000112 ***
MTD	0.7373	0.1368	5.389	1.73e-07 ***
COO	0.4684	0.2970	1.577	0.116161

Signif. codes: 0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1.

TABLE 4 50 random subsets of predictors of size 10 with $E_1 < 0.2$ and the corresponding E_2

	Predictors	E_1	E_2
1	MTD + BMRN+EHC + ITGR+BMY + PRGO+GKOS+ABBV+IDXX+ABT	0.1671	0.1224
2	EW + ARAV+MTD + PCRX+ICUI+CTMX+GILD+AMN + PRSC+CLVS	0.1718	0.1335
3	TAK + EW + MTD + MMSI+PFE + ANGO+MCK + PBH + AGN + NVO	0.0819	0.1781
4	MGLN+PCRX+CELG+GTS + MRTX+TAK + MTD + HQY + PBH + EW	0.1381	0.1906
5	INGN+HNGR+BDX + VIVO+AXGT+ABT + SPPI+AMGN+INOV+MTD	0.0895	0.1661
6	WMGI+MTD + SNN + ENSG+MGLN+NXGN+EW + INGN+CELG+XRAY	0.1982	0.1775
7	MTD + EW + UTHR+BIO + SGMO+CBMG+FGEN+TAK + ABT + OSUR	0.1513	0.1791
8	DVA + CBAY+MESO+WMGI+SNN + EW + AGN + PETS+MTD + THC	0.1732	0.155
9	MTD + PRSC+DBVT+INOV+CTMX+THC + BRKR+SRDX+STE + EW	0.1866	0.2001
10	TAK + COLL+RARE+MDRX+EPZM+BEAT+CDXS+INOV+NKTR+MDCO	0.1912	0.5035
11	VEEV+SNN + HRC + SGRY+BIO + LMNX+CBMG+THC + EW + MTD	0.1428	0.1654
12	QDEL+EPZM+SNN + CAH + MTD + OSUR+OFIX+ATRA+MRK + KPTI	0.1662	0.1752
13	HSTM+MTD + OSUR+MDCO+CAH + USPH+EW + UNH + SPPI+BMRN	0.1879	0.1844
14	UHS + HSIC+AZN + CNMD+BEAT+BAX + LMNX+PFE + ACHN+BLUE	0.1544	0.161
15	XON + AMPH+CERN+ABT + AXGT+ACAD+CGC + HSTM+MTD + HZNP	0.1808	0.1658
16	EBS + ALXN+PBH + MTD + IVC + PBYI+LHCG+CBMG+BAX + ARAV	0.1682	0.194
17	USPH+SPPI+WST + ADUS+MYGN+MRK + ONCE+MTD + BDX + BDSI	0.1611	0.1069
18	HQY + SRDX+MTD + BAX + ANIK+AMGN+IDXX+THC + PEN+WCG	0.1896	0.1739
19	HOLX+PCRX+EW + LCI + CAH + MCK + MTD + IONS+LH + ACHN	0.1486	0.1416
20	HQY + DERM+MTD + WST + PETS+MYGN+AMN + PEN+TAK + MDT	0.1954	0.1482
21	MASI+COLL+SNY + ACHN+MTD + CBM + ASND+HSTM+OMI + MOH	0.177	0.1386
22	MTD + ADUS+ASMB+MDRX+EW + XRAY+HCSG+ANIK+ASND+PETS	0.1788	0.1431
23	HSTM+MTD + ILMN+RAD+EW + SEM + CERN+ABC + NVCR+USPH	0.1632	0.1853
24	TFX + BKD + EW + PKI + MCK + ACAD+LIVN+SEM + AIMT+PBH	0.1919	0.3002
25	DVA + ABMD+BLUE+ICUI+ALGN+MTD + SGRY+VIVO+NTUS+BAX	0.1741	0.1787
26	EW + VIVO+CMD + MRK + PRGO+IRWD+MTD + MGLN+AZN + NUVA	0.1949	0.1856
27	MRK + IVC + WST + CRY+MTD + DBVT+RARE+ACHN+RMD + QDEL	0.1862	0.136
28	SYNH+MRK + PODD+VIVO+MGLN+AVNS+PDCO+WST + MTD + COO	0.0886	0.0855
29	MRK + MTD + CAH + QDEL+COLL+CPSI+MRTX+BIO + NVRO+XON	0.1805	0.129
30	VIVO+XRAY+TVTY+BIIB+SIEN+SYNH+A + EW + HSIC+BEAT	0.1001	0.1404
31	PRAH+LCI + SIEN+BRKR+RARE+SGMO+PDCO+AGN + DHR + SPPI	0.1913	0.1461
32	MTD + BAX + ABC + AMAG+ICUI+PBH + IDXX+SNY + DVA + EHC	0.1606	0.1625
33	WST + ACAD+ICUI+PTCT+AMN + IART+MTD + OMI + XRAY+BDSI	0.1718	0.1672
34	QGEN+MYGN+A + PRSC+CAH + MDCO+AZN + LHCG+PRAH+WAT	0.1994	0.1943
35	KMDA+MTD + BMY + WST + GMAB+IART+CGC + CTMX+OFIX+ATRI	0.1459	0.135
36	BAX + HZNP+ALNY+CLVS+HMSY+WCG + NVO + MESO+MTD + MRK	0.1863	0.1575
37	PCRX+AGN + JAZZ+CLVS+INVA+CMD + SRDX+MTD + BAX + SRPT	0.124	0.1936
38	OMCL+SEM + NVO + RDY + CVS + MTD + AMED+ATEC+PCRX+BAX	0.1386	0.1583
39	IRWD+DHR + TECH+NKTR+MTD + PDCO+LMNX+ILMN+OFIX+CLVS	0.1436	0.1683
40	PDCO+CVS + CBMG+MTD + PODD+EW + CRL + PETS+VEEV+ACHN	0.0827	0.1742
41	OFIX+BIO + ABT + SYK + UTHR+NTUS+BMY + BEAT+IDXX+PBH	0.1973	0.1855
42	LMNX+UTHR+HNGR+AMN + PRGO+MTD + EBS + SNY + MCK + MGLN	0.195	0.4046
43	SNN + TAK + CRY+DHR + CDXS+HSIC+MTD + FGEN+CBAY+SPPI	0.1665	0.1422

TABLE 4 (Continued)

	Predictors	E_1	E_2
44	ABT + INGN+ITGR+ASND+MTD + NXGN+VEEV+BRKR+NUVA+GLPG	0.1885	0.1722
45	AXGT+FGEN+PFE + GILD+EPZM+UTHR+VEEV+ALNY+LLY + BAX	0.1657	0.2008
46	DERM+GTS + MESO+EW + INVA+BEAT+USPH+CNMD+MYL + MTD	0.1928	0.1583
47	PACB+FGEN+BEAT+IVC + OFIX+MTD + MRTX+PBYI+ABT + OMI	0.1675	0.1649
48	IART+CBMG+PRSC+PRGO+STE + MTD + DBVT+EW + AMGN+PRAH	0.1963	0.17
49	RMD + AXGT+ABT + RAD+AMN + SRPT+USPH+PEN+MTD + IDXX	0.1817	0.1849
50	PDCO+MTD + EW + GLPG+RARE+VEEV+EHC + ABT + AMPH+FMS	0.1636	0.1267

Residual SE: 16.72 on 234 degrees of freedom.

Multiple R-squared: 0.6789, Adjusted R-squared: 0.6651.

F-statistic: 49.46 on 10 and 234 DF, *P*-value: < 2.2e-16.

The *P*-value is very small. Hence, there is sufficient evidence at the significance level of .01 to conclude that the regression on $X_1, X_2, ..., X_{10}$ explains a significant proportion of the variability in SP500.

Next we run the step-wise selection on the full model to find a simpler model.

```
step(full.lm)
```

As a result, we obtain the following reduced set of predictors: *SYNH*, *MRK*, *MGLN*, *AVNS*, *PDCO*, *WST*, *MTD*, and *COO*, named $X_1, X_2, ..., X_8$, respectively.

We are interested in testing for the significance of the collection of the two removed predictors in the step-wise selection. For this, we conduct the multiple partial *F* test. The null and alternative hypotheses are.

- $\mathbf{H_0}$: *SP500 ~ SYNH + MRK + MGLN + AVNS + PDCO + WST + MTD + COO* is the better model.
- $\mathbf{H_A}$: *SP500 ~ SYNH + MRK + PODD + VIVO + MGLN + AVNS + PDCO + WST + MTD + COO* is the better model.

Run the following R codes:

```
reduced = "SP500 ~ SYNH + MRK + MGLN
          + AVNS + PDCO + WST + MTD + COO"

reduced.lm = lm(reduced, data = df[1 : 245, ])

anova(reduced.lm, full.lm)
```

The output is as follows:

Model 1: SP500 ~ SYNH + MRK + MGLN + AVNS + PDCO + WST + MTD + COO

Model 2: SP500 ~ SYNH + MRK + PODD + VIVO + MGLN + AVNS + PDCO + WST + MTD + COO

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	236	65 456				
2	234	65 407	2	49.414	0.0884	0.9154

Since the *P*-value is .9154, we fail to reject the null hypothesis at the significance level of .01. In other words, there is insufficient evidence to conclude that the full model is better. Hence, we will use the reduced model since it is the simpler one.

3.5 | Check the multiple linear regression assumptions

In this section, we will check the multiple linear regression assumptions for the reduced model.

Multi-collinearity. We will investigate how the predictors are related to one another by computing the variance inflation factor (VIF) for each predictor.

Run the following R code:

```
car :: vif(reduced.lm)
```

The output is as follows:

SYNH	MRK	MGLN	AVNS	PDCO	WST	MTD	COO
1.378366	1.460910	1.202639	1.291041	1.135568	1.526046	2.004218	1.681826

Each predictor has VIF value much lower than 4. Hence, there is no evidence of serious multicollinearity.

Next we will check the four multiple linear regression assumptions for the reduced model.

Normality. Given any fixed combination of the predictors in the reduced model, the distribution of SP500 needs to be normally distributed.

Run the following R codes:

```
res = resid(reduced.lm)

qqnorm(res)

qqline(res)
```

The result is shown in Figure 1 below.

The normal Q-Q plot shows the sample quantiles and theoretical quantiles are highly correlated. Therefore, the normality assumption is satisfied.

Independence. The values of *SP500* must be independent, that is, form a random sample. This can be tested by verifying if the residuals from a linear model are correlated or not. To do this, we use the Durbin Watson test. The null and alternative hypotheses are.

H_0: There is no correlation among residuals, that is, they are independent.

H_A: The residuals are auto-correlated.

Run the following R code:

car::durbinWatsonTest(reduced.lm).

The result is indicated below:

lag	Autocorrelation	D-W Statistic	*P*-value
1	−0.07843169	2.155297	0.214
Alternative hypothesis: rho! = 0			

Since the *P*-value is .214, we fail to reject the null hypothesis at the significance level of .01. There is insufficient evidence to conclude that the residuals are auto-correlated. The test statistic *D-W* values in the range of 1.5 to 2.5 are relatively normal. The values under 1 or more than 3 are a definite cause for concern (see Field [2]). Since *DW* is approximately 2.2, there is no evidence against the independence assumption.

Linearity. The mean value of SP500 is a linear function of X_1, X_2, ..., X_8. In other words, the true statistical model is

$$E[\text{SP500}|X_1,X_2,...,X_8] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_8 X_8.$$

We will examine $E(\varepsilon)$, which needs to be close to 0, through the residuals vs fitted values plot.

Run the following R code:

```
plot(reduced.lm)
```

The result is as follows:

The Residuals vs Fitted plot in Figure 2 indicates linearity seems to hold reasonably well since the solid red line (residuals vs fitted values) is close to the dashed line (residuals = 0).

Homoscedasticity: The variance of SP500 is the same for any combination of values of X_1, X_2, ..., X_8.

Run the following R code:

```
plot(res)
```

The result is presented in Figure 3.

There is no clear pattern in the residual plot in Figure 3. Hence, the constant variance assumption is also satisfied.

3.6 | Test the reduced model

In this section, we will calculate the prediction error E_2 for the reduced model.

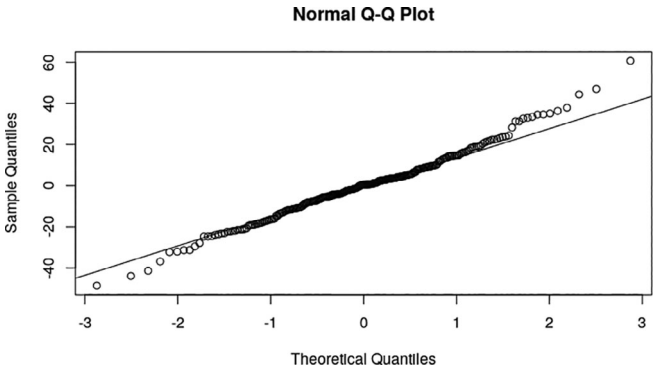

FIGURE 1 The normal Q-Q plot used to test for normality of the model

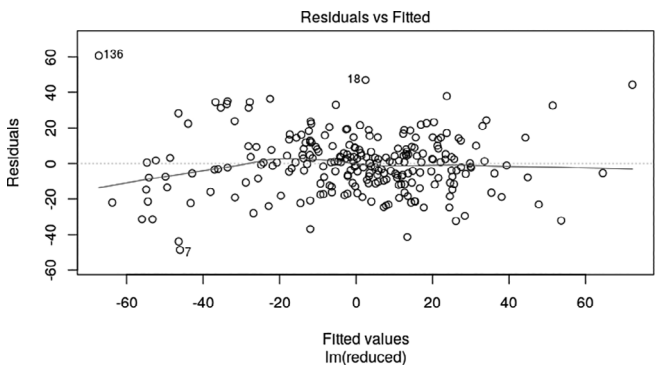

FIGURE 2 The Residuals vs Fitted plot used to test for linearity of the model

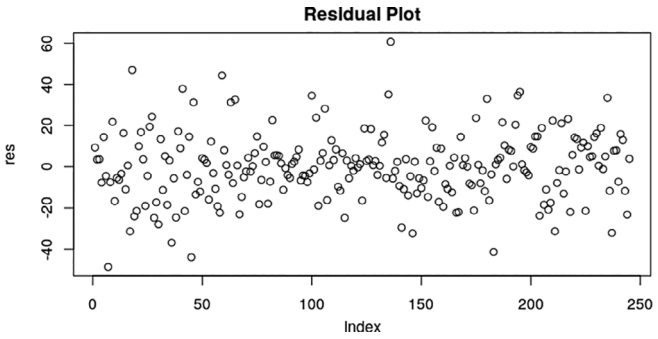

FIGURE 3 The residual plot used to test for homoscedasticity of the model

Run the following R code to summarize the reduced model.

```
summary(reduced.lm)
```

The simplified output is indicated below:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−0.09262	1.07725	−0.086	0.931559
SYNH	3.55223	0.92841	3.826	0.000167 ***
MRK	5.77396	1.34974	4.278	2.74e-05 ***
MGLN	1.74872	0.78105	2.239	0.026094 *
AVNS	2.29457	1.03944	2.208	0.028242 *
PDCO	11.78460	2.68370	4.391	1.70e-05 ***
WST	2.80812	0.69010	4.069	6.44e-05 ***
MTD	0.74263	0.13568	5.473	1.13e-07 ***
COO	0.47383	0.29556	1.603	0.110236

Signif. codes: 0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ “ 1.

Residual SE: 16.65 on 236 degrees of freedom.

Multiple R-squared: 0.6786, Adjusted R-squared: 0.6677.

F-statistic: 62.29 on 8 and 236 DF, *P*-value: < 2.2e-16.

We find the final model as follows:

$$SP500 = -0.09262 + (3.55223)SYNH + (5.77396)MRK + (1.74872)MGLN + (2.29457)AVNS + (11.78460)PDCO + (2.80812)WST + (0.74263)MTD + (0.47383)COO.$$

The R-squared is 0.6786. Hence, approximately 68% of the variability in SP500 is explained by the regression model.

Finally, we will test the effectiveness of the reduced model. We will compare the predicted values and the observed values at rows 251, 252, 253, 254, and 255 (Table 5).

The average absolute percentage error E_2 is 0.0813 or 8.13%.

4 | CONCLUSION

In multiple linear regression, models with higher R-squared do not necessarily have better predictability. In this example of health care stocks, for R-squared of at least 0.7, we are 99% confident that the prediction error is between 33.8% and 35.24%, which is high. The paper also emphasizes the importance of learning the data behavior in the process of producing a “best” subset of predictors. The pattern obtained from learning the data indicates it is reasonable to use the near out-of-sample prediction to control the far out-of-sample prediction. In other words, by setting the near out-of-sample prediction error to be small, there is a better chance to find a subset of predictors which can be used to produce a model with high predictability. In particular, by setting the near out-of-sample error E_1 less than 0.1, we obtain 10 predictors: *SYNH*, *MRK*, *PODD*, *VIVO*, *MGLN*, *AVNS*, *PDCO*, *WST*, *MTD*, and *COO*, which give a model with the approximate prediction error of 0.0855, and R-squared of 0.6789. By running the step-wise selection method, we find a simpler model with eight predictors: *SYNH*, *MRK*, *MGLN*, *AVNS*, *PDCO*, *WST*, *MTD*, and *COO*. The predictability of the final model is improved slightly with the prediction error of 0.0813 or 8.13%, and R-squared of 0.6786. The final model has much higher predictability and smaller R-squared than those in Table 2. Moreover, there is no evidence of serious multi-collinearity and the multiple linear regression assumptions are satisfied in the final model.

TABLE 5 Compare the predicted returns and observed returns of *S&P 500* over the testing time period

Row	Observed *S&P 500* Return	Predicted *S&P 500* Return	Percentage Error
251	−36.49	−39.2528	−0.0757
252	−52.6399	−45.0913	0.1434
253	23.0198	22.9484	0.0031
254	41.3801	34.7332	0.1606
255	−13.22	−13.5342	−0.0238

ORCID

Phong Luu https://orcid.org/0000-0001-8905-6681

REFERENCES

1. D. Enke, M. Grauer, and N. Mehdiyev, *Stock market prediction with multiple regression, fuzzy type-2 clustering and neural networks*, Proc. Comput. Sci. 6 (2011), 201–206.
2. A. P. Field, *Discovering statistics using SPSS: and sex and drugs and rock 'n' roll*, 3rd ed., Sage, London, 2009.
3. Investopedia, available at https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp
4. Investopedia, available at https://www.investopedia.com/terms/s/sp500.asp
5. Magnifymoney, available at https://www.magnifymoney.com/blog/investing/large-cap-stocks-vs-small-cap/
6. National Association of Insurance Commissioners, U.S. Health Insurance Industry, 2018 Annual Results.
7. R. Seethalakshmi, *Analysis of stock market predictor variables using Linear Regression*, Int. J. Pure Appl. Math. 119(15) (2018), 369–378.
8. T. Smith and A. Hawkins, *An economic regression model to predict market movements*, Int. J. Math. Trends and Technology. 28 (2015), 1–3. https://doi.org/10.14445/22315373/IJMTT-V28P501.

How to cite this article: Kieu T, Luu P, Yoon N. Multiple linear regression: Identify potential health care stocks for investments using out-of-sample predictions. *Teaching Statistics*. 2020;1–10. https://doi.org/10.1111/test.12233