The Stata Journal (2022)
22, Number 4, pp. 941–957 DOI: 10.1177/1536867X221141021

portfolio: A command for conducting portfolio analysis in Stata

Hongbing Zhu
Hohai University
Nanjing, China
zhuhongbing@hhu.edu.cn

Lihua Yang
Hohai University
Nanjing, China
yanglihua@hhu.edu.cn

Abstract. Portfolio analysis is widely used in empirical asset pricing to explore the cross-sectional relation between two or more variables. In this article, we introduce the methodology of portfolio analysis and describe a new command, `portfolio`, that provides a one-step solution for portfolio analysis. `portfolio` calculates the equal- or value-weighted returns with a t statistic for the portfolio and tests the significance of a long-short strategy in portfolios. `portfolio` also provides the Newey–West standard-error adjustment option for alleviating the impact of potential autocorrelation and heteroskedasticity in financial time series.

Keywords: st0697, portfolio, portfolio analysis, nonparametric analysis, hedging strategy

1 Introduction

Portfolio analysis is one of the most widely used statistical methods in empirical asset pricing. It is used to explore cross-sectional pricing factors for equity assets. Numerous pricing factors have been well revealed in the cutting-edge studies of empirical asset pricing through portfolio analysis (see, for example, Fama and French [2015]; Lee et al. [2019]; Hendershott, Livdan, and Rösch [2020]). Portfolio analysis is a nonparametric approach and therefore does not require any assumptions prior to the analysis. In practice, portfolio analysis can also help quantitative researchers evaluate the return performance of investment strategies. Because diversification, portfolio selection, and hedging are fundamental components of the modern finance curriculum, the Statistical Software Component Archive provides several community-contributed commands for financial analysis. A partial list is `fetchportfolio` (Dicle 2013), `tftools` (Dicle and Levendis 2017), and `cntrade` (Zhang and Li 2014). However, the community-contributed commands for portfolio analysis are still not available in official Stata.

In this article, we introduce the `portfolio` command, which provides the solution for conducting univariate or bivariate portfolio analysis, where the bivariate analysis methods can be set to either independent or dependent. The `portfolio` command also provides the Newey–West (1987) standard-error adjustment option accounting for the potential autocorrelation and heteroskedasticity problems of the financial time series. To conclude, the `portfolio` command allows researchers to conduct empirical asset pricing studies more efficiently. It expands the library of commands in Stata about financial investments.

The rest of the article is organized as follows. Section 2 documents the methodology of portfolio analysis for the `portfolio` command. Section 3 discusses the `portfolio` syntax. Section 4 provides several examples illustrating the usage of the `portfolio` command. Section 5 concludes.

2 Methodology

2.1 Univariate portfolio analysis

In empirical asset pricing studies, univariate portfolio analysis is used to reveal the relationship between a variable and the stock's future returns. Let us denote the variables involved in this analysis as Y and X. Typically, Y is the outcome variable, such as the stock's future returns. X is the risk-pricing variable, such as the stock's capital asset pricing model beta risk, which is a measure of the systematic risk of a security or portfolio compared with the market as a whole (Sharpe 1964; Lintner 1965). A standard approach to implementing univariate portfolio analysis consists of three steps.

The first step is the calculation of periodic breakpoints based on the value of X, which are used to group the entities in the sample into portfolios. If the investor wants to form N_p asset portfolios in time period t, then at least $N_p - 1$ breakpoints need to be computed. We denote the kth breakpoint in time period t as $B_{t,k}$ ($k = 1, 2, \ldots, N_p - 1$). In many cutting-edge empirical asset pricing studies, $B_{t,k}$ is generally identified by the quantile of the variable X in time period t, for example, the quintiles and deciles of variable X. In our program, we use the same quantile approach to determine breakpoints. We follow Bali, Engle, and Murray (2016) and define the breakpoints as

$$B_{t,k} = \text{Percentile}_{p_k}(X_t)$$

where Percentile$(\cdot)$ is the quantile function and p_k denotes the percentile that determines the kth breakpoint; the kth breakpoint in time period t is calculated as the p_kth percentile of the values of X across all entities in the sample in time period t. All of these computations are based on the nonmissing values of variable X. The procedure in the first stage will generate a set of breakpoints at time period t, $\{B_{t,1}, B_{t,2}, \ldots, B_{t,N_p-1}\}$.

Having identified the breakpoints in each time period t, we next construct portfolios. Specifically, at each time period t, individuals are classified into the first portfolio if the value of X of these individuals is less than or equal to the first breakpoint, $B_{t,1}$. The second portfolio comprises entities for which the value of corresponding X is less than or equal to the second breakpoint, $B_{t,2}$, but greater than the first breakpoint, $B_{t,1}$, and so on. Finally, the N_pth portfolio includes the entities with value of X satisfying the condition of $X > B_{t,N_p-1}$. Overall, the procedure in the second stage will generate several portfolios, each of which contains entity i, satisfying

$$\text{Portfolio}_{t,k} = \{i | B_{t,k-1} < X_{i,t} < B_{t,k}\} \quad k = 1, 2, \ldots, N_p \tag{1}$$

where Portfolio$_{t,k}$ denotes the kth portfolio at time period t, which includes several individuals i. The values of $B_{t,0}$ and B_{t,N_p} are defined as $-\infty$ and ∞, respectively, to ensure the validity of (1). Furthermore, in the third step, the average value of the outcome variable Y for the portfolios formed in the second step is computed at each time period t. Thus, we have

$$\overline{Y}_{t,k} = \frac{\sum\limits_{i \in \text{Portfolio}_{t,k}} W_{i,t} Y_{i,t}}{\sum\limits_{i \in \text{Portfolio}_{t,k}} W_{i,t}}$$

where $\overline{Y}_{t,k}$ refers to the average value of the outcome variable Y for the kth portfolio at time period t. $W_{i,t}$ is the weight variable that accounts for the contribution of entity i to the average outcome of portfolio k in time period t. In other words, it allows the user to consider the heterogeneity of different individual outcomes in the univariate portfolio analysis. If the user assumes that all entities in the portfolio have the same contribution, he or she can set the value of $W_{i,t}$ equal to 1 for all entities in each time period t. However, the widely used weights in empirical asset pricing studies are equal weights and value weights. Typically, value weights imply an allocation of investment amounts based on the market capitalization of the stock as a percentage of the portfolio. Nevertheless, the procedure in this stage will generate a set of time series, $\{\overline{Y}_{t,1}, \overline{Y}_{t,2}, \ldots, \overline{Y}_{t,N_p}\}$.

A range of statistical analyses can be conducted based on these time series. For example, when the outcome variable Y denotes the stock's returns excluding risk-free returns in time period $t+1$, one might be curious about whether the X-based portfolio earns an excess return. In this case, the user can calculate the average value of the Y for the portfolio k and conduct a statistical test with the null hypothesis: $\overline{Y}_k = 0$. Moreover, if one is interested in the significance of a zero-cost long-short strategy, for example, the significance of the return gained by going long portfolio 1 and going short portfolio N_p at time period t, one can first calculate the sequence of the difference between the excess returns of the two portfolios: $\text{dif}_t = \overline{Y}_{t,1} - \overline{Y}_{t,N_p}$. Then the same statistical test with the null hypothesis $\overline{\text{dif}_t} = 0$ can be conducted for checking the validity of the L/S strategy. In summary, the univariate portfolio analysis is expected to produce the following results in table 1.

Table 1. A brief summary of the univariate portfolio analysis

The kth portfolio	Average of outcome	Statistics for testing null hypothesis: $\overline{Y}_k = 0$
1	$\overline{Y}_1 = \frac{1}{T}\sum_{t\in T}\overline{Y}_{t,1}$	$t = \left(\overline{Y}_1 - 0\right)/\text{se}\left(\overline{Y}_1\right)$
2	$\overline{Y}_2 = \frac{1}{T}\sum_{t\in T}\overline{Y}_{t,2}$	$t = \left(\overline{Y}_2 - 0\right)/\text{se}\left(\overline{Y}_2\right)$
...	...	...
N_p	$\overline{Y}_{N_p} = \frac{1}{T}\sum_{t\in T}\overline{Y}_{t,N_p}$	$t = \left(\overline{Y}_{N_p} - 0\right)/\text{se}\left(\overline{Y}_{N_p}\right)$
L/S strategy: going long portfolio 1 and going short portfolio N_p	$\overline{\text{dif}} = \frac{1}{T}\sum_{t\in T}(\overline{Y}_{t,1} - \overline{Y}_{t,N_p})$	$t = \left(\overline{\text{dif}} - 0\right)/\text{se}\left(\overline{\text{dif}}\right)$

In addition to the above-mentioned calculation of average portfolio returns using the average arithmetic method, one can also obtain average alpha returns by regressing the portfolio returns ($\overline{Y}_{t,k}$) on several risk factors ($\widetilde{F}_m, m = 1, 2, \ldots, M$). The regression model can be expressed as

$$\overline{Y}_{t,k} = \alpha_k + \sum_{m\in M}\beta_m\widetilde{F}_m + \varepsilon_{t,k} \quad k = 1, 2, \ldots, N_p$$

Thus, the alpha returns adjusted by risk factors are

$$\widehat{\alpha}_k = E\left(\overline{Y}_{t,k} - \sum_{m\in M}\widehat{\beta_m}\widetilde{F}_m\right) \quad k = 1, 2, \ldots, N_p$$

Note that, because stock returns often exhibit autocorrelation and heteroskedasticity features, a t test based on the raw standard errors of the portfolio return series could be biased. To alleviate such bias, one can adjust the standard errors using the Newey–West (1987) approach. The adjusted standard errors take the form

$$\text{se} = \sqrt{\frac{\frac{1}{T}\left\{\sum_{t=1}^{T}e_{t,k}^2 + 2\sum_{l=1}^{L}\sum_{t=l+1}^{T}\left(1 - \frac{l}{1+L}\right)e_{t,k}e_{t-l,k}\right\}}{T}} \quad k = 1, 2, \ldots, N_p$$

where $e_{t,k}$ denotes the demeaned returns of the portfolio k: $e_{t,k} = \overline{Y}_{t,k} - E\left(\overline{Y}_{t,k}\right)$. The specific form of $E\left(\overline{Y}_{t,k}\right)$ depends on whether the average excess return of the portfolio is adjusted by the risk factor. $E\left(\overline{Y}_{t,k}\right)$ equals $\overline{Y}_k$ when the payoff of the portfolio is the average excess return; $E\left(\overline{Y}_{t,k}\right)$ equals $\widehat{\alpha}_k + \sum_{m \in M} \widehat{\beta}_m \widetilde{F}_m$ when the payoff of the portfolio is the risk-adjusted alpha return. Symbol L refers to the maximum lag length set by the user, and other variables are defined as the same in the previous equations.

2.2 Bivariate portfolio analysis

Bivariate portfolio analysis introduces an additional control variable to the univariate portfolio analysis that accounts for the effects of any other variables when examining the relationship between variables X and Y. Bivariate portfolio analysis also typically consists of three steps: computation of periodic breakpoints, formation of the portfolio, and examination of the portfolio's performance. The most salient differences between bivariate portfolio analysis and univariate portfolio analysis are the determination of the breakpoints and the formation of portfolios. Therefore, we focus on illustrating the processing method of these two aspects in bivariate portfolio analysis.

Let us denote the additional control variable as M. Then, we need to calculate the breakpoints of X and the breakpoints of M. There are two different techniques for calculating the breakpoints of X and M. The first is to compute the breakpoints of X and M independently as in the calculation of breakpoints in univariate portfolio analysis; this approach is often called an independent sort. The second is to begin by calculating breakpoints for the control variable M and then to compute breakpoints for X within each breakpoint interval of the control variable; this is known as a dependent sort. Nevertheless, these two approaches are essentially the same in calculating breakpoints for the control variable M. Specifically, we denote the jth breakpoint of the control variable M in time period t as $B^1_{t,j}$. Then $B^1_{t,j}$ satisfies

$$B^1_{t,j} = \text{Percentile}_{P1_j}(M_t) \quad j = 1, 2, \ldots, N_{P1} - 1$$

where Percentile$(\cdot)$ is the quantile function and $P1_j$ refers to the percentile that determines the jth breakpoint; the jth breakpoint in time period t is computed as the $P1_j$th percentile of the values of M across all entities in the sample in time period t. Next we should note that, under the first approach, the breakpoints for variable X are defined as

$$B^2_{t,k} = \text{Percentile}_{P2_k}(X_t) \quad k = 1, 2, \ldots, N_{P2} - 1 \tag{2}$$

where $B^2_{t,k}$ is the kth breakpoint of the core variable X in time period t. The definitions of any other elements are the same as in (2). Note that, in contrast to the definition of $B^2_{t,k}$ above, under the framework of dependent sorts, the breakpoints for variable X are defined as

$$B^2_{t,j,k} = \text{Percentile}_{P2_k}\left(X_t | B^1_{t,j-1} < M_t \leq B^1_{t,j}\right)$$

where $B^1_{t,0} = -\infty$ and $B^1_{t,N_{P1}} = \infty$. The definition of any other elements remains consistent. Equation (3) indicates that the breakpoints of variable X are conditional on the value of the control variable M. Having identified the breakpoints of the two variables at different time periods t, we next construct the portfolios. In bivariate portfolio analysis, the portfolio is defined as the set of common entities within the two breakpoint intervals. Thus, with the independent-sort approach, the portfolio is defined as

$$\text{Portfolio}_{t,j,k} = \{i|B^1_{t,j-1} < M_{i,t} \leq B^1_{t,j}\} \cap \{i|B^2_{t,k-1} < X_{i,t} \leq B^2_{t,k}\} \tag{3}$$

whereas, with the dependent-sort approach, the portfolio is defined as

$$\text{Portfolio}_{t,j,k} = \{i|B^1_{t,j-1} < M_{i,t} \leq B^1_{t,j}\} \cap \{i|B^2_{t,j,k-1} < X_{i,t} \leq B^2_{t,j,k}\} \tag{4}$$

Equation (3) [or (4)] will generate $j \times k$ portfolios, and one can further test the performance of these portfolios using the same approach as in univariate portfolio analysis. Equation (3) generates portfolios largely consistent with (4). And in practice, bivariate portfolio analysis based on dependent sorts is the most widely used.

3 The portfolio command

`portfolio` is a new community-contributed command for portfolio analysis described in section 2. It conducts several analyses commonly used in empirical asset pricing studies, including calculating equal- or value-weighted returns for the portfolio, testing the significance of a long-short strategy. `portfolio` allows one to use either a univariate portfolio test or a bivariate portfolio test, where the bivariate test methods can be either independent or dependent. `portfolio` also provides the Newey–West standard-error adjustment option, allowing users to alleviate the impact of potential autocorrelation and heteroskedasticity in the return series.

`portfolio` requires the following community-contributed commands: `astile` (Shah 2017b), `asgen` (Shah 2017a), `gduplicates` (Bravo 2018), and `spread` (Gomez 2021).

3.1 Syntax

`portfolio` has the following syntax.

`portfolio` *depvar* [*varlist_riskfactors*] [*if*] [*in*] `, groupby(`*timevar keyvar*`)`
`nq(`*#*`)` [`rf(`*var*`)` `w(`*var*`)` `cv(`*var*`)` `cvnq(`*#*`)` `independentsort` `lag(`*#*`)` `reg` `all`
`save(`*string*`)`]

3.2 Options

`groupby(`*timevar keyvar*`)` specifies the time frequency and the sort variable for portfolio construction. `groupby()` is required.

`nq(`*#*`)` specifies the number of quantiles. For example, `nq(10)` will create deciles, making 10 equal groups of the data based on the values of *keyvar*. `nq()` is required.

`rf(`*var*`)` specifies the risk-free rate variable for computing the portfolio's excess returns.

`w(`*var*`)` is the weights variable for the computation of portfolio returns.

cv(*var*) specifies the controlling variable, which accounts for the effects of any other variables when examining the relationship between the *keyvar* and the *depvar*.

cvnq(#) specifies the number of quantiles of the variable specified in the cv() option.

independentsort is valid only when the cv() option is specified. It specifies the construction method of the portfolio. The default construction method is the bivariate dependent sort, and one can override it to the bivariate independent sort by using this option.

lag(#) specifies the length of lags for estimating the Newey–West-consistent standard error.

reg saves the regression results of the portfolio analysis by the estimates store command. One can print the results by running the command esttab factorLoading* (Jann 2007).

all determines whether to print the detailed results of the portfolio test.

save(*string*) enables one to save the portfolios' outcome time series into a .dta file.

3.3 Stored results

portfolio stores the following in r():

Matrices

r(result)	main results of the portfolio analysis
r(tstat)	t statistic of the corresponding portfolio's return
r(p_val)	p-value corresponds to the t statistic

4 Examples

We use returndata.dta to illustrate the use of portfolio. returndata.dta contains stock transaction data for the Chinese A-share market from January 2017 to December 2018. The data in this dataset are available on several public finance platforms, such as the NetEase Finance platform and the East Money platform. In our example, we use a version of the data available at https://github.com/zhbsis/portfolio. The dataset has 10 variables and 78,078 observations. Details about the dataset are summarized as follows:

```
. use returndata

. describe

Contains data from returndata.dta
 Observations:        78,078
    Variables:            10                  30 Jan 2022 22:24
--------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------
stkcd           long    %06.0f                stock code
ym              float   %tm                   year_month
RET             float   %9.0g                 stock_returns:log(P{t}/P{t-1})
rf              double  %10.0g                risk-free rate
w               double  %10.0g                the circulation market value
mkt_rf          double  %10.0g                MKT factor
smb             double  %10.0g                SMB factor
hml             double  %10.0g                HML factor
L1_TO           float   %9.0g                 stock turnover in t-1 month
L1_PRICE        float   %9.0g                 stock log price in t-1 month
--------------------------------------------------------------------------------
Sorted by: stkcd  ym
```

In empirical asset pricing studies, researchers are usually concerned about the cross-sectional pricing factors of equity assets. For example, a stock turnover effect is well-documented by Zhang, Chen, and Yeh (2021). It reveals a significant negative relationship between stock turnover and the next month's stock returns. One can earn a significant excess return by going short in the portfolio with the highest (the largest quintile group) stock turnover and going long in the portfolio with the lowest (the smallest quintile group) stock turnover. The `portfolio` command is designed to examine the above relationships efficiently.

In practice, when the researcher is interested in the performance of a zero-cost long-short strategy that holds the bottom 20% of firms as ranked by monthly stock turnover and sells short the top 20%, the `portfolio` command can be conducted as follows:

```
. portfolio RET, groupby(ym L1_TO) nq(5) rf(rf)

(0 observations deleted)
(0 observations deleted)
        Univariate portfolio test with Huber and White t-statistic

-------------------------
         L1_TO |       ret
---------------+---------
             1 |  -0.01133
             2 |  -0.01565
             3 |  -0.01817
             4 |  -0.02207
             5 |  -0.03253
           5-1 |  -0.02120
         tstat |  -2.57323
         p_val |   0.01734
-------------------------

. matrix ew_res0 = r(result)
```

The results show that the average excess return of portfolio 1 equals −1.13% per month, while it equals 3.25% per month for portfolio 5. The L/S strategy by going long portfolio 5 and going short portfolio 1 generates an average excess return of −2.12% ($t = -2.57$) per month without considering transaction costs.

One can also compute the alpha return of the portfolio by including the risk factors in the `portfolio` command.

```
. portfolio RET mkt_rf smb hml, groupby(ym L1_TO) nq(5) rf(rf)
(0 observations deleted)
(0 observations deleted)
        Univariate portfolio test with Huber and White t-statistic
```

L1_TO	ret
1	0.00352
2	0.00521
3	0.00451
4	0.00200
5	-0.00731
5-1	-0.01082
tstat	-2.15643
p_val	0.04407

```
. matrix ew_res1 = r(result)
```

This command produces outputs in the same form, but the value in the table is the risk-adjusted alpha return of the corresponding portfolios. It shows that after considering the risk factors, the excess return (alpha return) by going long portfolio 5 and going short portfolio 1 increases to −1.08% ($t = -2.16$) per month. Furthermore, one can compute the value-weighted excess return of the portfolio by adding weights in option `w()`.

```
. portfolio RET, groupby(ym L1_TO) nq(5) rf(rf) w(w)
(0 observations deleted)
(0 observations deleted)
        Univariate portfolio test with Huber and White t-statistic
```

L1_TO	ret
1	0.00191
2	-0.00173
3	-0.00754
4	-0.01638
5	-0.02269
5-1	-0.02460
tstat	-2.34156
p_val	0.02867

```
. matrix ew_res2 = r(result)
```

```
. portfolio RET mkt_rf smb hml, groupby(ym L1_TO) nq(5) rf(rf) w(w)
(0 observations deleted)
(0 observations deleted)
        Univariate portfolio test with Huber and White t-statistic
```

L1_TO	ret
1	0.00438
2	0.01073
3	0.00690
4	0.00299
5	-0.00146
5-1	-0.00585
tstat	-0.99519
p_val	0.33215

```
. matrix ew_res3 = r(result)
```

It is also convenient to combine the above results with the `matlist` command:

```
. matrix onesorts = ew_res0, ew_res1, ew_res2, ew_res3
. matrix colname onesorts = ew:er ew:alpha vw:er vw:alpha
. matlist onesorts, showcoleq(c)
```

	ew		vw	
	er	alpha	er	alpha
1	-.0113272	.0035158	.0019073	.0043849
2	-.0156516	.0052114	-.0017271	.0107341
3	-.0181738	.0045055	-.0075398	.0069042
4	-.0220668	.0020006	-.0163777	.002988
5	-.03253	-.0073081	-.0226944	-.0014636
5-1	-.0212028	-.0108239	-.0246016	-.0058484
tstat	-2.573226	-2.156428	-2.341564	-.9951865
p_val	.0173424	.0440744	.0286657	.3321517

In addition, researchers may be more interested in the significance of portfolio returns or in the loadings of the risk factors because they capture the level of risk exposure of the portfolio returns. In that case, one can obtain it by adding the `reg` and `all` options. The saved factor loadings can be listed by the `esttab` command.

```
. portfolio RET mkt_rf smb hml, groupby(ym L1_TO) nq(5) rf(rf) w(w) reg all
(0 observations deleted)
(0 observations deleted)
Linear regression                               Number of obs     =         23
                                                F(3, 19)          =      25.44
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8210
                                                Root MSE          =     .01639

------------------------------------------------------------------------------
             |               Robust
   VW_Ri_Rf1 | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      mkt_rf |   .7086696   .0979076     7.24   0.000     .5037465    .9135926
         smb |   -.321777   .1454045    -2.21   0.039    -.6261121   -.0174419
         hml |   .3397191   .1009724     3.36   0.003     .1283815    .5510568
       _cons |   .0043849   .0036556     1.20   0.245    -.0032663    .0120361
------------------------------------------------------------------------------

Linear regression                               Number of obs     =         23
                                                F(3, 19)          =     140.46
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9635
                                                Root MSE          =     .00819

------------------------------------------------------------------------------
             |               Robust
   VW_Ri_Rf2 | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      mkt_rf |    1.00162   .0516545    19.39   0.000     .8935064    1.109734
         smb |   .0911009   .0637619     1.43   0.169    -.0423543    .2245562
         hml |   .0476866   .0341746     1.40   0.179    -.0238416    .1192149
       _cons |   .0107341   .0019075     5.63   0.000     .0067416    .0147265
------------------------------------------------------------------------------

Linear regression                               Number of obs     =         23
                                                F(3, 19)          =     108.56
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9474
                                                Root MSE          =     .01126

------------------------------------------------------------------------------
             |               Robust
   VW_Ri_Rf3 | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      mkt_rf |   1.131702   .0656972    17.23   0.000     .9941964    1.269208
         smb |   .1151142   .0888571     1.30   0.211     -.070866    .3010943
         hml |   .0189918   .0565397     0.34   0.741    -.0993471    .1373307
       _cons |   .0069042   .0023959     2.88   0.010     .0018895     .011919
------------------------------------------------------------------------------
```

```
Linear regression                               Number of obs     =         23
                                                F(3, 19)          =      83.73
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9248
                                                Root MSE          =     .01417

------------------------------------------------------------------------------
             |               Robust
   VW_Ri_Rf4 | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      mkt_rf |   1.150378   .0903584    12.73   0.000     .9612556      1.3395
         smb |   .5108827   .0991015     5.16   0.000      .303461    .7183045
         hml |   -.019552   .0832752    -0.23   0.817    -.1938491     .154745
       _cons |    .002988   .0028827     1.04   0.313    -.0030456    .0090215
------------------------------------------------------------------------------

Linear regression                               Number of obs     =         23
                                                F(3, 19)          =      46.16
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8505
                                                Root MSE          =     .02029

------------------------------------------------------------------------------
             |               Robust
   VW_Ri_Rf5 | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      mkt_rf |   .9940674   .1166282     8.52   0.000     .7499617    1.238173
         smb |   .7330654   .1539704     4.76   0.000     .4108016    1.055329
         hml |  -.2035167   .1338347    -1.52   0.145     -.483636    .0766026
       _cons |  -.0014636   .0046575    -0.31   0.757    -.0112118    .0082847
------------------------------------------------------------------------------

Linear regression                               Number of obs     =         23
                                                F(3, 19)          =      21.15
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7141
                                                Root MSE          =     .02899

------------------------------------------------------------------------------
             |               Robust
    __000001 | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      mkt_rf |   .2853979   .1869107     1.53   0.143    -.1058107    .6766064
         smb |   1.054842     .24127     4.37   0.000     .5498584    1.559826
         hml |  -.5432359   .1808344    -3.00   0.007    -.9217266   -.1647451
       _cons |  -.0058484   .0058767    -1.00   0.332    -.0181486    .0064517
------------------------------------------------------------------------------

        Univariate portfolio test with Huber and White t-statistic

-----------------------
       L1_TO |      ret
-------------+---------
           1 |  0.00438
           2 |  0.01073
           3 |  0.00690
           4 |  0.00299
           5 | -0.00146
         5-1 | -0.00585
       tstat | -0.99519
       p_val |  0.33215
-----------------------
```

```
        Detailed results of portfolio test

        Group |        ret       tstat       p_val
--------------+-------------------------------------
            1 |    0.00438     1.19950     0.24508
            2 |    0.01073     5.62730     0.00002
            3 |    0.00690     2.88165     0.00956
            4 |    0.00299     1.03651     0.31297
            5 |   -0.00146    -0.31424     0.75676

Factor loadings are stored and can be displaied by 'esttab factorLoading*'
. esttab factorLoading*, nogap varlabels(_cons alpha) compress varwidth(5)
```

	(1) VW_Ri_Rf1	(2) VW_Ri_Rf2	(3) VW_Ri_Rf3	(4) VW_Ri_Rf4	(5) VW_Ri_Rf5	(6) __000001
mkt_f	0.709***	1.002***	1.132***	1.150***	0.994***	0.285
	(7.24)	(19.39)	(17.23)	(12.73)	(8.52)	(1.53)
smb	-0.322*	0.0911	0.115	0.511***	0.733***	1.055***
	(-2.21)	(1.43)	(1.30)	(5.16)	(4.76)	(4.37)
hml	0.340**	0.0477	0.0190	-0.0196	-0.204	-0.543**
	(3.36)	(1.40)	(0.34)	(-0.23)	(-1.52)	(-3.00)
alpha	0.00438	0.0107***	0.00690**	0.00299	-0.00146	-0.00585
	(1.20)	(5.63)	(2.88)	(1.04)	(-0.31)	(-1.00)
N	23	23	23	23	23	23

```
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

If the researcher is concerned that the stock turnover effect is influenced by stock prices, then the `cv()` and `cvnq()` options can be added to the `portfolio` command for bivariate portfolio analysis. And the Newey–West approach can be used by adding the `lag()` option to adjust the standard errors of the portfolio's abnormal return.

```
. portfolio RET, groupby(ym L1_TO) nq(5) cv(L1_PRICE) cvnq(3) rf(rf) lag(2)
(0 observations deleted)
(0 observations deleted)
        Bivariate portfolio test with Newey-West t-statistic
```

L1_TO	L1_PRICE 1	2	3
1	0.00900	-0.02042	-0.02297
2	-0.01511	-0.01844	-0.01408
3	-0.01674	-0.01880	-0.01750
4	-0.01967	-0.02438	-0.02390
5	-0.02839	-0.03357	-0.03441
5-1	-0.03739	-0.01315	-0.01144
tstat	-3.27466	-1.50634	-1.66133
p_val	0.00346	0.14620	0.11083

The results show that the turnover effect gradually disappears as the stock price of the portfolio increases, and the excess return of the L/S strategy based on stock turnover is not statistically significant in column 3 (portfolio with the highest average stock price). For example, in the portfolio with the lowest stock price, the L/S strategy by going long portfolio 5 and going short portfolio 1 generates an excess return of −3.74% ($t = -3.27$) per month, while in the portfolio with the highest stock price, the excess return generated by the same L/S strategy increases to −1.14% ($t = -1.66$) per month.

If the user cares only about whether the relationship between the stock turnover and the next month's return is significant after controlling for the stock price, then a simple version of the result can be obtained by running the following command:

```
. astile G1 = L1_PRICE, nq(3) by(ym)
. astile G2 = L1_TO, nq(5) by(G1 ym)
. portfolio RET, groupby(ym G2) nq(5) rf(rf) lag(2)
(0 observations deleted)
(0 observations deleted)
        Univariate portfolio test with Newey-West t-statistic
```

G2	ret
1	-0.01146
2	-0.01588
3	-0.01768
4	-0.02265
5	-0.03212
5-1	-0.02066
tstat	-2.55284
p_val	0.01814

Finally, the `portfolio` command allows the user to store the original return series of the portfolio by adding the `save()` option. The `portfolio` command also stores some of the main results, which can be displayed with the following command:

```
. portfolio RET, groupby(ym L1_TO) nq(5) rf(rf) save("savedata")
(0 observations deleted)
(0 observations deleted)
        Univariate portfolio test with Huber and White t-statistic
```

L1_TO	ret
1	-0.01133
2	-0.01565
3	-0.01817
4	-0.02207
5	-0.03253
5-1	-0.02120
tstat	-2.57323
p_val	0.01734

```
file savedata.dta saved
Portfolio's return series are saved in savedata.dta
```

```
. return list
matrices:
             r(result) :  8 x 1
              r(tstat) :  5 x 1
              r(p_val) :  5 x 1
```

5 Conclusion

Although portfolio analysis is widely used in empirical asset pricing studies, there are no associated community-contributed solution commands in the Statistical Software Component Archives. In this article, we presented a new command, `portfolio`, that provides a one-step solution for portfolio analysis. Our command enables the investor to construct the portfolio based on a core variable and calculate the abnormal return of the portfolio and its t statistic. Besides, the command allows users to perform bivariate portfolio analysis by adding control variables. To cope with the potential autocorrelation and heteroskedasticity effects of financial time series, `portfolio` also provides users with the Newey–West standard-error adjustment option. In summary, `portfolio` enables researchers to conduct empirical asset pricing studies more efficiently. It also enriches the community-contributed commands on financial investments in Stata.

6 Acknowledgments

We express our sincere thanks to the editors and anonymous reviewers for their constructive comments and suggestions. This work is supported by the Graduate Research and Innovation Projects of Jiangsu Province (grant number KYCX21_0441) and the Fundamental Research Funds for the Central Universities (grant number B210201043 & B220203024).

7 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 22-4
. net install st0697      (to install program files, if available)
. net get st0697          (to install ancillary files, if available)
```

8 References

Bali, T. G., R. F. Engle, and S. Murray. 2016. *Empirical Asset Pricing: The Cross Section of Stock Returns*. Hoboken, NJ: Wiley.

Bravo, M. C. 2018. gtools: Stata module to provide a fast implementation of common group commands. Statistical Software Components S458514, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458514.html.

Dicle, M. F. 2013. Financial portfolio selection using the multifactor capital asset pricing model and imported options data. *Stata Journal* 13: 603–617. https://doi.org/10.1177/1536867X1301300310.

Dicle, M. F., and J. D. Levendis. 2017. Technical financial analysis tools for Stata. *Stata Journal* 17: 736–747. https://doi.org/10.1177/1536867X1701700312.

Fama, E. F., and K. R. French. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116: 1–22. https://doi.org/10.1016/j.jfineco.2014.10.010.

Gomez, M. 2021. stata-tidy. GitHub. https://github.com/matthieugomez/tidy.

Hendershott, T., D. Livdan, and D. Rösch. 2020. Asset pricing: A tale of night and day. *Journal of Financial Economics* 138: 635–662. https://doi.org/10.1016/j.jfineco.2020.06.006.

Jann, B. 2007. Making regression tables simplified. *Stata Journal* 7: 227–244. https://doi.org/10.1177/1536867X0700700207.

Lee, C. M. C., S. T. Sun, R. Wang, and R. Zhang. 2019. Technological links and predictable returns. *Journal of Financial Economics* 132: 76–96. https://doi.org/10.1016/j.jfineco.2018.11.008.

Lintner, J. 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47: 13–37. https://doi.org/10.2307/1924119.

Newey, W. K., and K. D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55: 703–708. https://doi.org/10.2307/1913610.

Shah, A. 2017a. asgen: Stata module to generate weighted average mean using an existing variable as weight. Statistical Software Components S458409, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458409.html.

———. 2017b. astile: Stata module to provide a faster and byable alternative for xtile. Statistical Software Components S458321, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s458321.html.

Sharpe, W. F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x.

Zhang, B., W. Chen, and C.-Y. Yeh. 2021. Turnover premia in China's stock markets. *Pacific-Basin Finance Journal* 65: 101487. https://doi.org/10.1016/j.pacfin.2020.101487.

Zhang, X., and C. Li. 2014. Importing Chinese historical stock market quotations from NetEase. *Stata Journal* 14: 381–388. https://doi.org/10.1177/1536867X1401400209.

About the authors

Hongbing Zhu is a lecturer in finance at Hohai University in Nanjing, China.

Lihua Yang is a PhD candidate in management science and engineering at Hohai University, China. She is the corresponding author.