The Unit-Weighted Mean - Because Size Matters*

March 30, 2020

Eugene Canjels†

Abstract

The unit-weighted mean is of frequent interest to applied researchers in a wide range of fields. Despite this interest, there is a lack of easily accessible theoretical statistical literature that shows its statistical properties. This paper provides the asymptotic distribution of the unit-weighted mean and a formula to calculate asymptotically valid standard errors. I show that numerically identical results can be obtained using a novel regression approach.

1 Introduction

Applied researchers are frequently interested in an outcome per unit. Some examples are crop yield (production per acre), concentration of a chemical in a solution (moles per liter), or investment returns (profit per dollar invested). When the observations collected by the researcher vary by the number of units in each observation, the statistical measure of interest is often the unit-weighted mean. For instance, if the observation unit is a farm, the researcher observes total output and total acres per farm - and thus crop yield - for each farm. The researcher can then calculate total output over all farms divided by total

*The Securities and Exchange Commission disclaims responsibility for any private publication or statement of any SEC employee or Commissioner. This article expresses the author's views and does not necessarily reflect those of the Commission, the Commissioners, or other members of the staff. This paper is part of the Division of Economic and Risk Analysis' Working Paper Series. Papers in this series are the work of the authors and not the work of the Division or the Commission. Inclusion of a paper in this series does not indicate a Division or Commission determination to take any particular action or position. References to this paper should indicate that the paper is a "DERA Working Paper."

†Division of Econonomic and Risk Analysis, U.S. Securities and Exchange Commission, 100 F St NE, Washington DC 20549, canjelse@sec.gov, Phone: 202-551-8515.

acres over all farms, or equivalently calculate the acre-weighted average yield. Introducing notation, the unit-weighted mean $\widetilde{r}$ is calculated as:

$$\widetilde{r} = \frac{\sum y_i}{\sum u_i} = \frac{\sum u_i r_i}{\sum u_i} = \sum \frac{u_i}{\sum u_i} r_i = \sum w_i r_i \tag{1.1}$$

where $r_i = y_i/u_i$ and $w_i = u_i/\sum u_i$.[1] In general, the unit-weighted mean and the unweighted mean are measuring different variables of interest. For example, when measuring crop yield, the unit-weighted mean measures the productivity of a region while the unweighted mean measures the productivity of an average farm in that region. These measures will differ when farm size and productivity are correlated.

Statistical comparisons are the natural extension to the calculation of the unit-weighted mean. Is the crop yield in region A different from the crop yield in region B, and is this difference statistically significant? The existing statistical literature on weighted means focusses on three types of weights: frequency weights, probability weights, and precision weights. The unit-weighted mean requires a different analysis which is discussed in this paper.

Much of the analysis in this paper is not novel, and statisticians familiar with sampling literature may recognize some of the central results. First, some results in this paper are discussed in the sampling literature under the term "ratio estimation," e.g. Cochran (1977). In that literature, the focus tends to be on sampling approaches to estimate $\sum y_i$ in a finite population using u_i as an auxiliary variate. Second, the problem analyzed here can be interpreted as a cluster sampling problem. In that interpretation the "acre" is the fundamental element of the analysis, and the "farm" is a cluster of acres. The analysis here is thus similar to single-stage cluster sampling, meaning that clusters are randomly drawn from a super-population and all elements in the cluster are sampled. The applicability of the sampling literature to the analysis of the unit-weighted mean appears to be underappreciated by empirical researchers. And by abstracting from the issues of estimating means in a finite population, the statistical theory here is easier to follow and more accessible to applied researchers.

I extend the existing literature in a couple of ways. First, I show that the estimation error results from two components, with the total variance approximately the sum of the variance of the two components. The first error results from drawing a set of observations u_i that may not be representive of the dis-

[1] Summations in this paper always run from 1 to n.

tribution of sizes in the population. The second error results from drawing observations y_i that may not be representative of the expected value of y_i given the observed values of u_i.

I also introduce a regression approach that provides numerically identical results. The ratio estimation literature shows that the weighted mean is an optimal approach to estimate $\sum y_i$ when y_i and u_i have a linear relation going through the origin and a variance that is proportional to u_i. These assumptions are likely to be overly restrictive for most applied work. But I show that the weighted mean can be calculated using this regression approach and that the Huber-White approach for heteroscedasticity provides the correct standard errors.

2 Asymptotic distribution of the unit-weighted mean

Let the n pairs (u_1, y_1), (u_2, y_2), ..., (u_n, y_n) be drawn independently and with equal probability from a super-population of pairs, i.e. the pairs (u_i, y_i) are i.i.d. with probability distribution $g(u_i, y_i)$. We require that the number of units u_i is strictly positive. We already defined $r_i = y_i/u_i$ earlier. The parameters μ_u, μ_y, and μ_r indicate the population means for u_i, y_i, and r_i, respectively. Similarly, the population variances are given by σ_u^2, σ_y^2, and σ_r^2, and population covariances given by σ_{uy}, σ_{ur}, and σ_{yr}. We assume that the probability distribution $g(u_i, y_i)$ satisfies the necessary conditions for the Lindeberg-Levy Central Limit Theorem (i.e. finite variances and covariances). Sample analog estimators for the population parameters are given the usual notation, e.g. $\bar{u} = n^{-1}\sum u_i$, $s_u^2 = n^{-1}\sum (u_i - \bar{u})^2$, and $s_{uy} = n^{-1}\sum (u_i - \bar{u})(y_i - \bar{y})$.[2]

The following straightforward result provides the probability limit of the unit-weighted mean.

Theorem 1. Let $\mu_{\widetilde{r}} = \mu_y/\mu_u$. *The weighted mean* $\widetilde{r}$ *is a consistent estimator of* $\mu_{\widetilde{r}}$.

Proof.

$$\text{plim}\ \widetilde{r} = \text{plim}\ \frac{\frac{1}{n}\sum y_i}{\frac{1}{n}\sum u_i} = \frac{\mu_y}{\mu_u} = \mu_{\widetilde{r}}$$

□

Note however that, in general, $\widetilde{r}$ is not an unbiased estimator because

[2] All results below are easily adjusted if one prefers to use unbiased sample variances and covariances that use $n-1$ instead of n in the denominator.

$$E\left(\sum y_i / \sum u_i\right) \neq E\left(\sum y_i\right) / E\left(\sum u_i\right)$$

.

The unit-weighted mean and the unweighted mean measure different population parameters with the difference between the two parameters given by covariance between u_i and r_i divided by the expectation of u_i as shown in the following derivation:

$$\begin{aligned}
\mu_{\widetilde{r}} - \mu_r &= \frac{\mu_y - \mu_u \mu_r}{\mu_u} \\
&= \frac{E\left(u \cdot r\right) - \mu_u \mu_r}{\mu_u} \\
&= \frac{E\left(\left(u - \mu_u\right)\left(r - \mu_r\right)\right)}{\mu_u} \\
&= \frac{\sigma_{ur}}{\mu_u} \qquad (2.1)
\end{aligned}$$

Thus the unit-weighted and the unweighted mean only coincide when the covariance between u_i and r_i is equal to zero. Similarly, focusing on the sample analogs, the difference between $\widetilde{r}$ and $\bar{r}$ is the sample covariance between the scaled up weights $(n \cdot w_i)$ and r_i:

$$\begin{aligned}
cov\left(n \cdot w_i, r_i\right) &= \frac{1}{n} \sum \left(n w_i - n \bar{w}\right)\left(r_i - \bar{r}\right) \\
&= \frac{1}{n} \sum n w_i r_i - n \bar{w} \bar{r} \\
&= \widetilde{r} - \bar{r} \qquad (2.2)
\end{aligned}$$

Given that the unit-weighted mean is a type of averaging, one may expect that its asymptotic distribution is normal. This is correct as shown in the following theorem:

Theorem 2. *The asymptotic distribution of* $\widetilde{r}$ *is given by:*[3]

$$\widetilde{r} \overset{a}{\sim} N\left(\mu_{\widetilde{r}}, \phi^2 / n\right)$$

[3] To make notation easier to read we use $\overset{a}{\sim}$ as short-hand for "is approximately distributed as" instead of the more formal notation where the difference between the random variable and its limiting value, scaled by the standard error, converges to the standard normal.

where

$$\phi^2 = \left(\frac{1}{\mu_u}\right)^2 \left(\sigma_y^2 + \mu_{\widetilde{r}}^2 \sigma_u^2 - 2\mu_{\widetilde{r}}\sigma_{uy}\right)$$

Proof. The multivariate version of the Lindeberg-Levy CLT shows that the joint distribution of $\bar{u}$ and $\bar{y}$ is asymptotically normal. Following the delta-method, the Taylor series expansion of the non-linear function $\widetilde{r} = \bar{y}/\bar{u}$ is given by

$$\widetilde{r} \overset{a}{=} \mu_{\widetilde{r}} + \frac{1}{\mu_u}(\bar{y} - \mu_y) - \frac{\mu_y}{\mu_u^2}(\bar{u} - \mu_u) = \mu_{\widetilde{r}} + \frac{1}{\mu_u}\left\{(\bar{y} - \mu_y) - \mu_{\widetilde{r}}(\bar{u} - \mu_u)\right\}$$

Asymptotic normality and the formula for the variance follow directly. □

3 Estimation of the standard error of the unit-weighted mean.

Because ϕ depends on unknown population parameters, it will need to be estimated to make Theorem 2 useful in practice. This leads to the following result:

Theorem 3. *Under the assumptions above*

$$\widetilde{r} \overset{a}{\sim} N\left(\mu_{\widetilde{r}}, s_{\widetilde{r}}^2\right)$$

where

$$s_{\widetilde{r}}^2 = \sum w_i^2 (r_i - \widetilde{r})^2$$

Proof. Let the estimatorf^2 be the sample analog of ϕ^2, i.e.

$$f^2 = \left(\frac{1}{\bar{u}}\right)^2 \left(s_y^2 + \widetilde{r}^2 s_u^2 - 2\widetilde{r}s_{uy}\right)$$

then

$$\text{plim } f^2 = \phi^2$$

The second term in parentheses of the estimator f^2 simplifies:

$$
\begin{aligned}
& s_y^2 + \widetilde{r}^2 s_u^2 - 2\widetilde{r} s_{uy} \\
&= \frac{1}{n}\sum\left(y_i - \bar{y}\right)^2 + \frac{1}{n}\widetilde{r}^2\sum\left(u_i - \bar{u}\right)^2 - \frac{2}{n}\widetilde{r}\sum\left(y_i - \bar{y}\right)\left(u_i - \bar{u}\right) \\
&= \frac{1}{n}\sum y_i^2 - \bar{y}^2 + \frac{1}{n}\sum \widetilde{r}^2 u_i^2 - \widetilde{r}^2\bar{u}^2 - \frac{2}{n}\sum \widetilde{r} y_i u_i + 2\widetilde{r}\bar{y}\bar{u} \\
&= \frac{1}{n}\sum u_i^2 r_i^2 + \frac{1}{n}\sum u_i^2 \widetilde{r}^2 - \frac{2}{n}\sum u_i^2 r_i \widetilde{r} - \bar{y}^2 - \left(\frac{\bar{y}}{\bar{u}}\right)^2 \bar{u}^2 + 2\left(\frac{\bar{y}}{\bar{u}}\right)\bar{y}\bar{u} \\
&= \frac{1}{n}\sum u_i^2\left(r_i - \widetilde{r}\right)^2
\end{aligned}
$$

such that

$$
f^2 = \left(\frac{1}{\bar{u}}\right)^2 \frac{1}{n}\sum u_i^2\left(r_i - \widetilde{r}\right)^2 = n\sum w_i^2\left(r_i - \widetilde{r}\right)^2
$$

Using the Slutsky Theorem, we can replace ϕ in Theorem 2, with its (consistent) estimator f, finishing the proof. □

4 Conditioning on the weights

It is tempting to analyze the statistical properties of $\tilde{r}$ taking u as fixed and writing the following equation for the variance of $\tilde{r}$:

$$
\begin{aligned}
E\left(\left[\tilde{r} - E\left(\tilde{r}|u_i\right)\right]^2 |u_i\right) &= E\left(\left[\sum w_i r_i - w_i E\left(r_i|u_i\right)\right]^2 |u_i\right) \\
&= \sum w_i^2 E\left(\left[r_i - E\left(r_i|u_i\right)\right]^2 |u_i\right) \qquad (4.1)
\end{aligned}
$$

It would then be tempting to estimate $E\left(\left[r_i - E\left(r_i|u_i\right)\right]^2\right)$ using $\left(r_i - \widetilde{r}\right)^2$. Unfortunately, this approach does not work because $E\left(r_i|u_i\right)$ can not be simply replaced by $\widetilde{r}$.

To analyze the conditional distribution of $\widetilde{r}$, we define $E\left(r_i|u_i\right) = \mu_{r|u_i} = \mu_i$ and $V\left(r_i|u_i\right) = \sigma^2_{r|u_i} = \sigma_i^2$. We replace the tedious subscripts $r|u_i$ with just the subscript i since in this section r_i is the only source of uncertainty. We allow for the most general case where both the conditional mean and variance of r_i may vary by i.

Theorem 4. *Conditional on the values of u_i, and assuming that μ_i and σ_i^2 satisfy the Lindeberg-Feller conditions, the asymptotic distribution of $\widetilde{r}$ is given* *by*

$$\widetilde{r}|u \overset{a}{\sim} N\left(\mu_{\widetilde{r}|u}, \phi_w\right)$$

where $\mu_{\widetilde{r}|u} = \sum w_i \mu_i$ and $\phi_w = \sum w_i^2 \sigma_i^2$

Proof. The expectation is

$$E\left(\widetilde{r}|u\right) = E\left(\sum w_i r_i|u\right) = \sum w_i E\left(r_i|u\right) = \sum w_i \mu_i = \mu_{\widetilde{r}|u}$$

The variance is

$$\begin{aligned} V\left(\widetilde{r}|u\right) &= E\left(\left(\widetilde{r} - \mu_{\widetilde{r}|u}\right)^2 |u\right) = E\left(\left[\sum w_i \left(r_i - \mu_i\right)\right]^2 |u\right) \\ &= \sum w_i^2 E\left(\left(r_i - \mu_i\right)^2 |u\right) = \sum w_i^2 \sigma_i^2 = \phi_w \end{aligned}$$

The Lindeberg-Feller Central Limit Theorem finishes the proof. □

We take the estimator of the variance as before:

$$s_{\tilde{r}}^2 = \sum w_i^2 \left(r_i - \widetilde{r}\right)^2$$

The expectation of this estimator conditional on u is given by the following theorem.

Theorem 5. *Conditional on u, the expectation of $s_{\tilde{r}}^2$ is given by*

$$E\left(s_{\tilde{r}}^2|u\right) = \phi_w - 2\sum w_i^3 \sigma_i^2 + \sum w_i^2 \sum w_i^2 \sigma_i^2 + \sum w_i^2 \left(\mu_i - \mu_{\widetilde{r}|u}\right)^2$$

Proof. See Appendix □

In this formula, the first and the last terms are $O\left(n^{-1}\right)$, while the two middle terms are $O\left(n^{-2}\right)$. Therefore:

$$E\left(s_{\tilde{r}}^2|u\right) \overset{a}{=} V\left(\widetilde{r}|u\right) + \sum w_i^2 \left(\mu_i - \mu_{\widetilde{r}|u}\right)^2 \tag{4.2}$$

The term $\sum w_i^2 \left(\mu_i - \mu_{\widetilde{r}|u}\right)^2$ is an estimate of the variance of $\tilde{r}$ resulting from uncertainty in the drawing a set of observations with a distribution of units that is different from that of the population, and $V\left(\widetilde{r}|u\right)$ is the variance of

$\tilde{r}$ resulting from the uncertainty in drawing values of r_i that are different from the expectation of r_i given u_i.

Thus we can decompose the estimation error $\widetilde{r} - \mu_{\widetilde{r}}$ into two components:

$$\widetilde{r} - \mu_{\widetilde{r}} = \left(\widetilde{r} - \mu_{\widetilde{r}|u}\right) + \left(\mu_{\widetilde{r}|u} - \mu_{\widetilde{r}}\right) \tag{4.3}$$

and the total variance of the estimation error is approximately the sum of the variance of these two components.

In conclusion, when we think of $\widetilde{r}$ as an estimator of $\mu_{\widetilde{r}}$ we can think of the estimation error resulting from two sources. First, there is the error resulting from the values of u_i that were drawn that may not be representive of the population. Second, given the observed values of u_i, there is the error resulting from the values of y_i given the observed values of u_i. Those values of y_i may not be representative of the expected value of y_i given u_i. Without additional constraints on the problem, the decomposition is only of theoretical interest. The values μ_i are not known and it is not possible to construct an unbiased estimate of the variance conditional on the observed units. The data simply can't fully distinguish between observed returns coming from μ_i or uncertainty of the r_i given u_i.

5 Regression Equivalence

Consider running the following regression

$$\sqrt{u_i} r_i = \beta \sqrt{u_i} + \epsilon_i \tag{5.1}$$

The OLS estimate of β is then given by

$$\widehat{\beta} = \frac{\sum \sqrt{u_i}\sqrt{u_i} r_i}{\sum \left(\sqrt{u_i}\right)^2} = \frac{\sum u_i r_i}{\sum u_i} = \tilde{r} \tag{5.2}$$

Thus the unit-weighted mean can be calculated using the regression above, in an analogous way to calculating the simple mean by running a regression of r_i on a constant. One way to think about this is as a methods of moments estimator. The coefficient β is that number that solves the following equation:

$$E\left(\sqrt{u_i} \cdot \epsilon_i\right) = E\left(\sqrt{u_i} \cdot \left[\sqrt{u_i} r_i - \beta\sqrt{u_i}\right]\right) = E\left(u_i \cdot r_i - \beta u_i\right) = 0$$

such that

$$E\left(y_i\right) - \beta E\left(u_i\right) = 0$$

or

$$\beta = \frac{E\left(y_i\right)}{E\left(u_i\right)}$$

That is, $\widehat{\beta}$ estimates the ratio of the expectations of y_i and u_i as before.

The regression above clearly should not be interpreted as a causal relation, nor as a conditional expectation of $\sqrt{u_i}r_i$ given $\sqrt{u_i}$, but simply as a best linear predictor. That is, $\widehat{\beta}$ minimizes the squared prediction errors summed up over all the units:

$$\min_\beta \sum \epsilon_i^2 = \min_\beta \sum \left(\sqrt{u_i}r_i - \beta\sqrt{u_i}\right)^2 = \min_\beta \sum u_i \left(r_i - \beta\right)^2 \tag{5.3}$$

Given our assumption that all observations are equal-probability draws from a super-population, the error term ϵ_i is independent across observations, but it is not necessarily homoscedastic. This suggests estimating the standard error on $\widehat{\beta}$ using Huber-White heteroscedasticity robust standard errors. Those standand errors simplify in this situation exactly to those in Theorem 3:

$$\begin{aligned} v_{HCE}\left(\widehat{\beta}_{OLS}\right) &= \left(X'X\right)\left(X'diag\left(\widehat{\epsilon}_1^2, ..., \widehat{\epsilon}_n^2\right)X\right)\left(X'X\right)^{-1} \\ &= \left(\sum u_i\right)^{-1}\left(\sum\left(\sqrt{u_i}\right)^2\widehat{\epsilon}_i^2\right)\left(\sum u_i\right)^{-1} \\ &= \left(\sum u_i\right)^{-2}\left(\sum u_i\left(\sqrt{u_i}r_i - \widehat{\beta}\sqrt{u_i}\right)^2\right) \\ &= \frac{\sum u_i^2\left(r_i - \widehat{\beta}\right)^2}{\left(\sum u_i\right)^2} = \sum w_i^2\left(r_i - \tilde{r}\right)^2 \end{aligned} \tag{5.4}$$

Thus, the analysis of the unit-weighted mean can be done in standard statistical software as a simple regression with the standand errors estimated using the Huber-White technique.

This idea can be extended to test for equivalence of the weighted mean across two populations. That is, let

$$\widetilde{r}_g \overset{a}{\sim} N\left(\widetilde{\mu}_g, s_g^2\right)$$

for groups $g = A$ and $g = B$, and the observations in groups A and B are independent. The hypothesis $\widetilde{\mu}_A = \widetilde{\mu}_B$ can be tested using the statistic

$$\frac{\widetilde{r}_A - \widetilde{r}_B}{\sqrt{s_A^2 + s_B^2}}$$

Using the regression approach, it is straightforward to show that the same result is obtained as the t-statistic on γ in the regression

$$\sqrt{u_i} r_i = \beta \sqrt{u_i} + \gamma \sqrt{u_i} D_i + \epsilon_i \tag{5.5}$$

where D_i is a dummy variable indicating membership in group A or B.

6 Two Examples in Finance

6.1 IPO underpricing

When companies go public, shares are sold to the public in an Initial Public Offering or IPO. It is well known that the offering price is on average well below the closing price at the end of the first day of trading, a phenomenon referred to as IPO underpricing.[4] We will not discuss here the various reasons proposed in the literature for this empirical result, but show how the techniques shown above apply in this setting. We take data from Bloomberg Finance L.P. on IPOs from 1995 to 2019, including effective date, offer size, offer price, first-day returns (that is, return to first close). The results are shown Table 1. By year, we calculate the unweighted first-day returns and their standard errors using standard formulas, and the weighted first-day returns and their standard errors using Equation 1.1 and Theorem 3.

Investors who are considering investing in an IPO may be more interested in unweighted returns as they are more likely to provide information on what they may expect for the the next IPO. Researchers and regulators are more likely interested in economy wide effects and thus more likely to care about the proceeds weighted returns. The differences are not large for most years.[5] Over the entire 25 year period, the proceeds-weighted average return is 20.2% versus 23.7% for the unweighted average return showing that underpricing tends to be slightly smaller for larger IPOs.

[4] For instance, Ljungqvist, 2007.

[5] The difference in 2008 results from the IPO of Visa, Inc. Visa, Inc. was the single largest IPO in the 20 year period, and accounted for 78% of the entire offering proceeds in 2008. The Visa Inc. IPO was underpriced by 28.4%.

The results show large underpricing in all years, but especially around the dot-com boom of 1999-2000, with weighted average first-day returns of 63.1% and 44.8%. The standard errors for these years show that the IPOs in those years were fundamentally different from other years, and the returns did not just arise from normal randomness in IPO returns. The period from 2009 to 2012 following the great recession shows slightly lower first-day returns than the most recent period from 2013 to 2019. The differences, however, are small compared to the standard errors and its not clear whether this reflects a fundamental change or is just normal random variation.

6.2 Earnings-to-Price Ratios

The price-to-earnings ratio ("P/E Ratio") is a common metric used by financial analysts to evaluate the price of a stock. We pulled from Bloomberg Finance L.P. for all S&P 500 companies information as of December 31, 2019. This information includes the following variables: price per share, earnings per share, shares outstanding and primary exchange name. We deleted companies with no earnings information, and used A Class shares prices for companies with multiple share classes. This leaves us with 497 firms - 131 listed on Nasdaq, 365 listed on NYSE, and 1 listed on CBOE.

As an initial matter, the commonly used P/E ratio is a poor definition to use in a statistical analysis. The P/E ratio is not defined for zero earnings, behaves poorly for small earnings, and the meaning of the slope for negative earnings is different than for positive earnings (i.e. more negative P/E ratios are better than less negative P/E ratio for a constant value of P). We will therefore flip the ratio and investigate the earnings-to-price ratio (E/P ratio) which is better suited for a statistical analysis. Of course, in aggregate, the ratio of total market value to total earnings is the inverse of the ratio of total earnings to total market value. For this set of companies, with these variable definitions, the P/E ratio on December 31, 2019, was 21.5, equivalent to an E/P ratio of 4.7%. That is, a dollar of market value for S&P 500 companies was supported by 4.7 cents of earnings.

Table 2 shows the results of the analysis. Panel A uses the usual formulas for the unweighted average E/P ratios. In Panel B the top three lines calculate market cap weighted average P/E ratios and their standard errors using Equation 1.1 and Theorem 3. The last line of Panel A was generate using a regression with an intercept and a dummy for NYSE listed companies. The last line of

Panel B was generated using the same dummy variable and Equation 5.5.

In general, larger companies (in terms of market cap) are supported by fewer earnings. The unweighted mean across the 497 companies is 5.2% of earnings per dollar of market value, whereas the market cap weighted mean is 4.7%. The share price of companies on Nasdaq is generally supported by fewer earnings than those on New York, with a difference of 1.0% unweighted and 1.3% market cap weighted. These differences are statistically significant with t-statistics of 3.24 and 3.53 respectively.

References

[1] Cochran, W. 1977. *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons.

[2] Ljungqvist, A. 2007. IPO Underpricing. In *Handbook of Empirical Corporate Finance, Volume 1*, ed. E. Eckbo, 375-422. Amsterdam: Elsevier

Appendix

Proof Theorem 5.

Proof. Using the definitions and rules of expectations we get:

$$\begin{aligned}
E\left(s_{\widetilde{r}}^2\right) &= E\left(\sum w_i^2 \left(r_i - \widetilde{r}\right)^2 |u\right) \\
&= E\left(\sum w_i^2 \left(r_i - \sum w_j r_j\right)^2 |u\right) \\
&= E\left(\sum w_i^2 \left(\left[r_i - \mu_i\right] - \left[\sum w_j r_j - \mu_i\right]\right)^2 |u\right) \\
&= E\left(\sum w_i^2 \left(\left[r_i - \mu_i\right]^2 - 2\left[r_i - \mu_i\right]\left[\sum w_j r_j - \mu_i\right] + \left[\sum w_j r_j - \mu_i\right]^2\right) |u\right) \\
&= \sum w_i^2 E\left(\left[r_i - \mu_i\right]^2 |u\right) - 2\sum w_i^2 E\left(\left[r_i - \mu_i\right]\left[\sum w_j r_j - \mu_i\right] |u\right) \\
&+ \sum w_i^2 E\left(\left[\sum w_j r_j - \mu_i\right]^2 |u\right)
\end{aligned}$$

Evaluating the three expectations in this formula:

$$E\left(\left[r_i - \mu_i\right]^2 |u\right) = \sigma_i^2$$

$$E\left([r_i-\mu_i]\left[\sum w_j r_j-\mu_i\right]|u\right)=E\left([r_i-\mu_i]\left[\sum w_j\left(r_j-\mu_i\right)\right]|u\right)$$
$$=E\left([r_i-\mu_i]\left[w_i\left(r_i-\mu_i\right)\right]|u\right)$$
$$=E\left(w_i\left[r_i-\mu_i\right]^2|u\right)=w_i\sigma_i^2$$

and

$$\begin{aligned}
E\left(\left[\sum w_j r_j-\mu_i\right]^2|u\right) &= E\left(\left[\left(\sum w_j r_j-\sum w_j\mu_j\right)-\left(\mu_i-\sum w_j\mu_j\right)\right]^2|u\right)\\
&= E\left(\left(\sum w_j r_j-\sum w_j\mu_j\right)^2|u\right)\\
&- 2E\left[\left(\sum w_j r_j-\sum w_j\mu_j\right)\left(\mu_i-\sum w_j\mu_j\right)|u\right]\\
&+ E\left[\left(\mu_i-\sum w_j\mu_j\right)^2|u\right]\\
&= E\left\{\left(\sum w_j\left[r_j-\mu_j\right]\right)^2|u\right\}\\
&- 2E\left\{\left(\sum w_j\left[r_j-\mu_j\right]\right)\left(\mu_i-\sum w_j\mu_j\right)|u\right\}\\
&+ \left(\mu_j-\mu_{\widetilde{r}|u}\right)^2\\
&= \sum w_j^2\sigma_j^2+\left(\mu_j-\mu_{\widetilde{r}|u}\right)^2
\end{aligned}$$

Thus

$$\begin{aligned}
E\left(s_{\tilde{r}}^2|u\right) &= \sum w_i^2\sigma_i^2-2\sum w_i^3\sigma_i^2+\sum w_i^2\left[\sum w_j^2\sigma_j^2+\left(\mu_j-\mu_{\widetilde{r}|u}\right)^2\right]\\
&= \sum w_i^2\sigma_i^2-2\sum w_i^3\sigma_i^2+\sum w_i^2\sum w_i^2\sigma_i^2+\sum w_i^2\left(\mu_j-\mu_{\widetilde{r}|u}\right)^2
\end{aligned}$$

□

Table 1. IPO Underpricing, 1995 -2019

Data from Bloomberg Finance L.P. using function "IPO." IPO effective date between January 1, 1995, and December 31, 2019. Shares listed as either Common Stock or Class A Shares, and "US" as the country code on the Bloomberg symbol. Offering price of $5 or more. IPOs with missing or invalid first day returns are removed.

			First-day Return			
			Unweighted		Proceeds Weighted	
Year	**Number of IPOs**	**Aggregate Proceeds ($ Billions)**	**Mean**	**Std Error**	**Mean**	**Std Error**
1995	521	$27.07	20.4%	1.2%	17.8%	1.5%
1996	743	$40.56	16.3%	1.0%	16.5%	0.9%
1997	497	$30.19	15.4%	0.9%	17.5%	1.3%
1998	350	$34.97	20.9%	2.2%	16.9%	2.3%
1999	484	$52.98	68.4%	4.4%	63.1%	7.0%
2000	349	$56.12	53.5%	4.1%	44.8%	8.8%
2001	77	$35.46	14.1%	1.9%	8.9%	2.8%
2002	71	$19.48	9.7%	1.9%	6.0%	2.9%
2003	75	$10.94	15.8%	2.2%	15.8%	3.1%
2004	185	$35.93	12.3%	1.2%	13.9%	2.3%
2005	167	$30.85	10.2%	1.3%	10.2%	1.7%
2006	148	$28.69	13.3%	1.8%	20.6%	5.6%
2007	145	$31.17	13.6%	1.8%	13.2%	3.4%
2008	24	$25.05	8.1%	5.0%	24.9%	3.6%
2009	40	$13.62	11.3%	2.7%	11.8%	2.7%
2010	98	$33.50	7.3%	1.4%	6.0%	1.5%
2011	83	$26.34	12.6%	2.4%	10.0%	2.6%
2012	100	$33.39	17.1%	2.4%	9.8%	4.8%
2013	156	$43.73	20.2%	2.4%	19.8%	3.5%
2014	199	$42.71	15.4%	2.0%	12.9%	2.0%
2015	126	$24.89	19.3%	3.3%	17.1%	3.5%
2016	81	$12.37	14.6%	2.8%	14.9%	3.6%
2017	124	$25.79	12.3%	2.0%	16.5%	4.6%
2018	141	$32.21	16.7%	2.3%	18.3%	3.6%
2019	116	$40.78	23.3%	3.4%	18.2%	6.2%
1995-2019	**5100**	**$788.79**	**23.7%**	**0.7%**	**20.2%**	**1.1%**

Table 2. Earning-to-Price Ratios S&P 500 Companies, December 2019

Data from Bloomberg Finance L.P. S&P 500 companies excluding those with missing earnings, and dedupped for companies with multiple share classes. Price "PX_LAST", Trailing 12M Diluted EPS From Continuing Operations “RR844”, Shares Outstanding “BS081”, and Primary Exchange Name “DS197”.

Panel A. Unweighted Average E/P Ratio

					95% Confidence Interval	
Primary Exchange	**Count**	**Mean**	**Std Error**	**t-stat**	**Lower Bound**	**Upper Bound**
Nasdaq	131	4.5%	0.3%	17.90	4.0%	5.0%
New York	365	5.5%	0.2%	33.26	5.2%	5.8%
Total	**497**	**5.2%**	**0.1%**	**37.47**	**5.0%**	**5.5%**
Difference Nasdaq - New York		*1.0%*	*0.3%*	*3.24*	*0.4%*	*1.6%*

Panel B. Market Cap Weighted Average E/P Ratio

					95% Confidence Interval	
Primary Exchange	**Count**	**Mean**	**Std Error**	**t-stat**	**Lower Bound**	**Upper Bound**
Nasdaq	131	3.9%	0.3%	12.79	3.3%	4.4%
New York	365	5.1%	0.2%	26.02	4.7%	5.5%
Total	**497**	**4.7%**	**0.2%**	**24.05**	**4.3%**	**5.0%**
Difference Nasdaq - New York		*1.3%*	*0.4%*	*3.53*	*0.6%*	*2.0%*