Lecture 6
Return Predictability
What drives stock market prices? A present-value decomposition and application of AR models

Lars A. Lochstoer

UCLA Anderson School of Management

Winter 2020

Outline

1. Stock Market Predictability
    - Forecasting regressions
    - The Dividend-Yield
    - Cross-equation Restrictions (the Present-Value restriction)

2. References

3. Appendix: Background on Optimal Forecasting

Stock return predictability

- Let R_{t+1} denote the simple return on the aggregate market, e.g. the CRSP-VW index.
- Let D_t denote aggregate dividends and $d_t = \log(D_t)$.
- The ratio D_t / P_t is called the **dividend yield** while P_t / D_t is called the **price-to-dividend** ratio

Stock return predictability

- A **forecasting regression** is a regression of an outcome at time $t+j$ (with $j>0$) using an predictor variable known at time t:

$$y_{t+j} = \alpha + \beta x_t + \varepsilon_{t+j}, \quad \text{for } t = 1, \ldots, T$$

Table: Return Predictability

Regression	slope	t-stat	HAC t-stat	R^2
$R_{t+1} = a + b(D/P)_t + \varepsilon_{t+1}$	3.498	[2.309]	[2.395]	0.062
$R_{t+1} - R_t^f = a + b(D/P)_t + \varepsilon_{t+1}$	3.933	[2.621]	[2.726]	0.078
$r_{t+1} = a_r + b_r(dp)_t + \varepsilon_{t+1}^r$	0.105	[1.989]	[2.075]	0.047
$\Delta d_{t+1} = a_d + b_d(dp)_t + \varepsilon_{t+1}^d$	0.008	0.203	0.185	0.001

Notes: Annual Data. Sample 1927-2009. R_{t+1} is the real return on the CRSP-VW index. r_{t+1} denotes logs of the real return. R_{t+1}^f denotes the return on the real risk-free.

Interpretation

- an increase in the dividend yield of 1 percentage point in deviation from its mean increases the expected real return by 3.49 percentage points (per annum).
- note: when returns are regressed on lagged persistent variables such as the dividend/yield, the disturbances are correlated with the regressor's innovation; this tends to create an upward bias in the case of dividend-yield regressions and is called **Stambaugh bias**; see Stambaugh (1999).
- **Stambaugh bias** implies that OLS coefficients are estimated to be too high.

Relation between Regressions

- note that the log dividend/yield in deviation from its mean is (to a first-order Taylor expansion) given by:

$$dp_t = D_t/P_t \ / \ (D/P)$$

where D/P is the (unconditional) average dividend/price ratio

- so we can state the return regression :

$$r_{t+1} = a_r + b_r \times dp_t + u_{t+1}$$

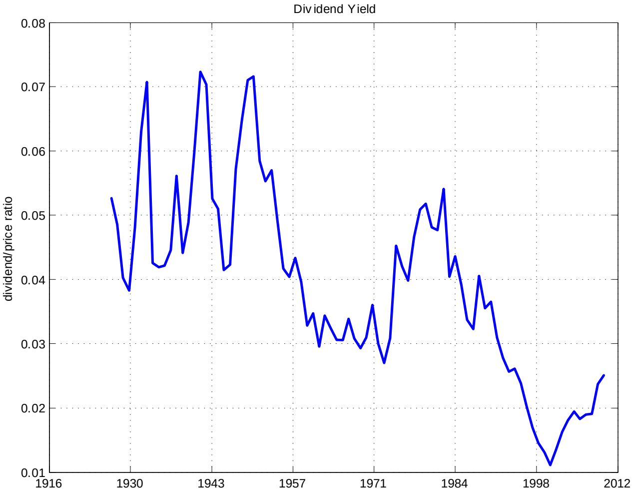
as follows:

$$r_{t+1} = a_r + b_r \times D_t/P_t \ / \ (D/P) + u_{t+1}$$

- the average dividend yield D/P is .035
- so the implied coefficient for the regression with the dividend yield is

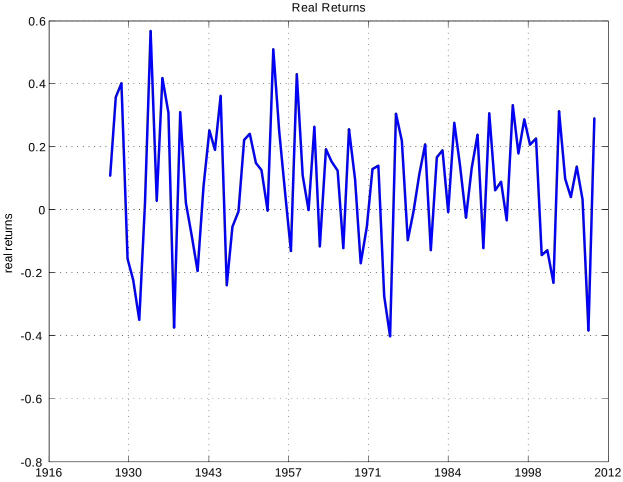
$$\frac{b_r}{D/P} = .105/.035 = 3.00$$

Dividend Yield



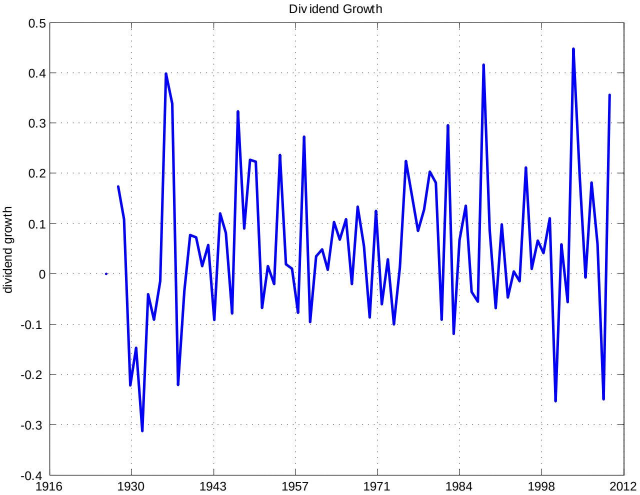
Dividend Yield on CRSP-VW (AMEX-NASDAQ-NYSE). Annual data. 1926-2009.

Real Returns



Real Returns on CRSP-VW (AMEX-NASDAQ-NYSE). Annual data. 1926-2009.

Dividend Growth



Dividend Growth on CRSP-VW (AMEX-NASDAQ-NYSE). Annual data. 1926-2009.

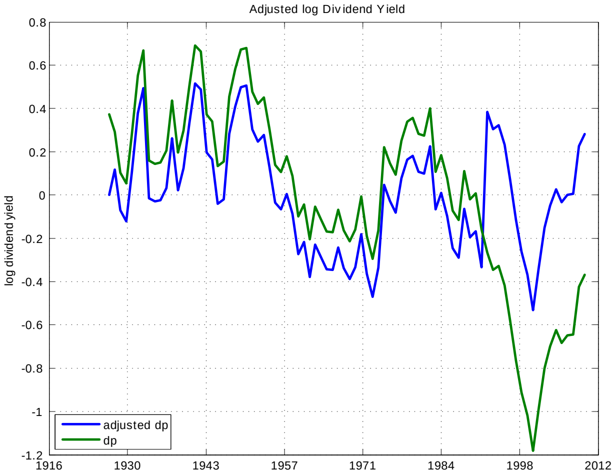
Structural Break in 1991

- Lettau and Van Nieuwerburgh (2007) find a structural break in log dividend yield in 1991.
- defined adjusted dividend yield:

$$
\begin{aligned}
\widetilde{dp}_t &= dp_t - \overline{dp}_1 \\
\widetilde{dp}_t &= dp_t - \overline{dp}_2
\end{aligned}
$$

where $\overline{dp}_1$ denotes the mean in the first sample 1926-1991 and $\overline{dp}_2$ denotes the mean in the second sample 1992-2009.

Log Dividend Yield



Demeaned log Dividend Yield $\widetilde{dp}_t$ on CRSP-VW (AMEX-NASDAQ-NYSE) with break in 1991. Annual data. 1926-2009.

Return Predictability

Table: Return Predictability

Regression	slope	t-stat	HC t-stat	R^2
$r_{t+1} = a_r + b_r(\widetilde{dp})_t + \varepsilon_{t+1}$	0.267	[3.118]	[3.667]	0.107
$\Delta d_{t+1} = a_d + b_d(\widetilde{dp})_t + \varepsilon_{t+1}$	0.039	[0.624]	[0.736]	0.004

Notes: Annual Data. Sample 1927-2009. R_{t+1} is the real return on the CRSP-VW index. R_{t+1} denotes logs of the real return. R^f_{t+1} denotes the return on the real risk-free.

Longer Horizons

- we run the following regression of k period holding returns on the dividend yield:

$$\sum_{i=1}^{k} r_{t+i} = a_r + b_r^k (dp)_t + \varepsilon_{t+k}$$

- as you increase the horizon k, the slope coefficients b_r^k increase and the R^2 increase

- Note: in this case, you should account for autocorrelation of residuals up to and including $k-1$ observations apart mechanically induced by the overlap
  - The next couple of slides shows how to do this using HAC standard errors

HAC robust standard errors

- If OLS residuals exihibit heteroskedasticity and/or autocorrelation (and, potentially, non-normality), OLS is still *consistent*
  - But, not efficient
  - Maximum likelihood is the efficient method in large samples
  - OLS is maximum likelihood only when errors are i.i.d. normally distributed

- If we still choose OLS (as a linear regression is pretty robust and parsimonious), we need to adjust the standard errors
  - HAC (heteroskedasticity and autocorrelation adjusted) standard errors

HAC robust standard errors: theory

- Please refer back to the "Note on Asymptotic Standard Errors" I posted earlier (which have already read)

- Recall, for the case of Asymptotic OLS

$$y_t = x_t\beta + \varepsilon_t, \quad \text{for } t = 1, ..., T$$

$$\hat{\beta}_T - \beta \overset{asymptotically}{\rightarrow} N\left(0, \frac{1}{T} E\left[x_t x_t'\right]^{-1} SE\left[x_t x_t'\right]^{-1}\right)$$

where

$$S = \sum_{j=-\infty}^{\infty} E\left[x_t x_{t-j}' \varepsilon_t \varepsilon_{t-j}\right]$$

and $\hat{\beta}_T$ is the estimate of β in a sample of length T

HAC robust standard errors: theory

- If the residuals are correlated across q leads and lags and zero thereafter

$$corr\left(\varepsilon_t, \varepsilon_{t-j}\right) \begin{cases} \neq 0 & \text{for } |j| \leq q \\ = 0 & \text{for } |j| > q \end{cases}$$

we have

$$S = \sum_{j=-q}^{q} E\left[x_t x_{t-j}' \varepsilon_t \varepsilon_{t-j}\right]$$

- These are called Hansen-Hodrick standard errors (see next slide)

Hansen-Hodrick standard errors

Define:

$$R_T(v;\beta) = \frac{1}{T}\sum_{t=1+v}^{T} x_t x_{t-v}' \varepsilon_t \varepsilon_{t-v}$$

where the estimate of the spectral density matrix is

$$\hat{S}_T = R_T\left(0;\hat{\beta}_T\right) + \sum_{v=1}^{q}\left[R_T\left(v;\hat{\beta}_T\right) + R_T\left(v;\hat{\beta}_T\right)'\right]$$

The estimate of the covariance matrix is then

$$Est.Asy.Var\left(\hat{\beta}_T\right) = T\left(X_T' X_T\right)^{-1}\hat{S}_T\left(X_T' X_T\right)^{-1}$$

where capital x_t, X_t, is a $T \times K$ matrix with t'th row equal to x_t

Newey-West standard errors

Newey and West (1987) solve an issue for the Hansen-Hodrick standard errors

- The estimated variance covariance matrix of $\hat{\beta}$ can be non-positive definite
- I.e., not invertible, "negative variance"
- To ensure a positive-definite covariance matrix, downweight estimated autocorrelations more the farther from the 0'th lag:

$$\hat{S}_T = R_T\left(0; \hat{\beta}_T\right) + \sum_{v=1}^{q} \frac{q+1-v}{q+1}\left[R_T\left(v; \hat{\beta}_T\right) + R_T\left(v; \hat{\beta}_T\right)'\right]$$

The Newey-West covariance matrix is then

$$Est.Asy.Var\left(\hat{\beta}\right) = T\left(X_T' X_T\right)^{-1} \hat{S}_T \left(X_T' X_T\right)^{-1}$$

- For Newey-West (NW) standard errors, should use $(k-1) \times 1.5$ or so due to the downweighting in the NW procedure
- Note that NW with 0 lags overlap is the same as White standard errors

Long-Horizon Return Predictability with Dividend Yield

Table: Return Predictability

Horizon	1	2	3	4	5
slope	0.105	0.199	0.250	0.282	0.323
OLS	[1.989]	[2.692]	[2.976]	[3.046]	[3.232]
NW	[2.036]	[2.399]	[2.578]	[2.573]	[2.600]
R^2	0.047	0.083	0.101	0.106	0.119

Notes: Annual Data. Sample 1927-2009. Forecasting regression of $\sum_{i=1}^{k} r_{t+i}$ on the log dividend yield. $\sum_{i=1}^{k} r_{t+i}$ denotes the sum of k years of logs of the real return.

Longer Horizons

- we run the following regression of k period holding returns on the dividend yield:

$$\sum_{i=1}^{k} r_{t+i} = a_r + b_r^k (\widetilde{dp})_t + \varepsilon_{t+k}$$

- as you increase the horizon k, the slope coefficients b_r^k increase and the R^2 increase
- Consider the 5 year horizon (next slide) where $\hat{b}_r^5 = 0.826$. An increase in the dividend yield of 1 percentage point in deviation from its mean increases the expected real return by 23.71 percentage points ($=.826/.035$) or 4.74 percentage points (per annum).

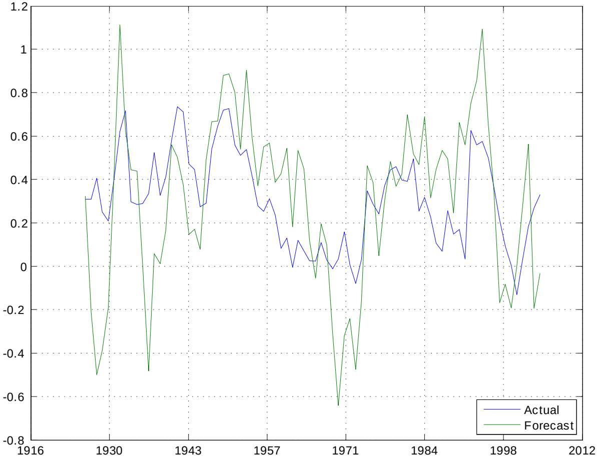
Long-Horizon Return Predictability with Adj. Div. Yield

Table: Return Predictability

Horizon	1	2	3	4	5
slope	0.267	0.478	0.661	0.750	0.826
OLS	[3.137]	[4.075]	[5.179]	[5.578]	[5.892]
NW	[3.480]	[3.960]	[4.977]	[4.559]	[4.157]
R^2	0.107	0.170	0.251	0.283	0.308

Notes: Annual Data. Sample 1927-2009. Forecasting regression of $\sum_{i=1}^{k} r_{t+i}$ on the adjusted log dividend yield. $\sum_{i=1}^{k} r_{t+i}$ denotes the sum of k years of logs of the real return.

5-year return Forecast



5-year log return forecast using Adjusted log Dividend Yield on CRSP-VW (AMEX-NASDAQ-NYSE). Annual data. 1926-2009.

Longer Horizons

- the R^2 in the regression of k period holding returns on the dividend yield is given by:

$$R^2(k) = \frac{V[E_t[r_{t+1}] + \ldots + E_t[r_{t+k}]]}{V[r_{t+1} + r_{t+2} + \ldots + r_{t+k}]}$$

- this grows at rate k initially because
  - realized returns are negatively autocorrelated
  - predicted returns are positively autocorrelated

Linearizing the returns

- consider the return on an asset:

$$R_{t+1} \equiv \frac{P_{t+1} + D_{t+1}}{P_t} = \frac{\frac{D_{t+1}}{D_t}(1 + PD_{t+1})}{PD_t}.$$

- pd_t denotes the log price-dividend ratio

$$pd_t = p_t - d_t = \log\left(\frac{P_t}{D_t}\right),$$

  where price is measured at the end of the period and the dividend flow is over the same period.
- also: note that

$$dp_t = -pd_t$$

Log-Linearizing returns

- Campbell and Shiller (1989) log-linearization of the return equation around the (unconditional) mean log price/dividend ratio delivers the following expression for log returns:

$$r_{t+1} = \Delta d_{t+1} + \rho pd_{t+1} + k - pd_t,$$

with linearization coefficients ρ and k that depend on the mean of the log price/dividend ratio $\overline{pd}$:

$$\rho = \frac{e^{\overline{pd}}}{e^{\overline{pd}} + 1} < 1 \quad \text{(the } k \text{ coefficient not important)}$$

- this expression is an approximation of an identity. It must hold!

The log of the price/dividend ratio

$$pd_t \;=\; \Delta d_{t+1} + \rho pd_{t+1} + k - r_{t+1}$$

- iterating forward on the linearized return equation
- imposing a no-bubble condition:

$$\lim_{j \to \infty} \rho^j pd_{t+j} = 0$$

- expression for the log price/dividend ratio:

$$pd_t = constant + \overbrace{\sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j}}^{\textbf{cash flow}} - \overbrace{\sum_{j=1}^{\infty} \rho^{j-1} r_{t+j}}^{\textbf{discount rate}}$$

Price/Dividend Ratios

Price/dividend ratios can only move if they predict returns or cash flows:

$$pd_t = constant + \overbrace{E_t \left[\sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j} \right]}^{\textbf{cash flow}} - \overbrace{E_t \left[\sum_{j=1}^{\infty} \rho^{j-1} r_{t+j} \right]}^{\textbf{discount rate}}$$

- a high price-to-dividend ratio pd_t implies that dividends are expected to increase or future returns (discount rates) are expected to decline

Campbell-Shiller Decomposition

The *pd* equation (without expectations) implies that the variance of the price/dividend ratio equals:

$$
\begin{aligned}
V[pd_t] &= V\left[\sum_{j=1}^{\infty}\rho^{j-1}\Delta d_{t+j}\right] + V\left[\sum_{j=1}^{\infty}\rho^{j-1}r_{t+j}\right] - 2\mathrm{cov}\left(\sum_{j=1}^{\infty}\rho^{j-1}r_{t+j}, \sum_{j=1}^{\infty}\rho^{j-1}\Delta d_{t+j}\right) \\
&= \mathrm{cov}\left(\sum_{j=1}^{\infty}\rho^{j-1}\Delta d_{t+j} - \sum_{j=1}^{\infty}\rho^{j-1}r_{t+j}, \sum_{j=1}^{\infty}\rho^{j-1}\Delta d_{t+j}\right) - \mathrm{cov}\left(\sum_{j=1}^{\infty}\rho^{j-1}\Delta d_{t+j} - \sum_{j=1}^{\infty}\rho^{j-1}r_{t+j}, \sum_{j=1}^{\infty}\rho^{j-1}r_{t+j}\right) \\
&= \mathrm{cov}\left(pd_t, \sum_{j=1}^{\infty}\rho^{j-1}\Delta d_{t+j}\right) - \mathrm{cov}\left(pd_t, \sum_{j=1}^{\infty}\rho^{j-1}r_{t+j}\right)
\end{aligned}
$$

- Campbell and Shiller: the price/dividend ratio has to predict future (long-run) returns and/or dividends if it moves around!
  - the evidence that it predicts returns seems stronger than the evidence that it predicts cash flows

Variance Decomposition

The variance decomposition of the log price/dividend ratio is the difference between two regression slope coefficients:

$$1 = \frac{\text{cov}\left(pd_t, \sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j}\right)}{V[pd_t]} - \frac{\text{cov}\left(pd_t, \sum_{j=1}^{\infty} \rho^{j-1} r_{t+j}\right)}{V[pd_t]}$$

- Is the variance of the *pd*-ratio driven by variation in expected cash flows or expected returns (i.e., discount rates)?

Price/Dividend Ratios

Price/dividend ratios predict future returns.

So do the term spread, the default spread and T-bill rates.

The R^2 increase with the forecasting horizon.

Variance Decomposition

- Recall that $dp_t = -pd_t$. Thus., the dp_t equation (without expectations) implies that the the slope coefficients in a regression of discounted returns and dividend growth on dp_t satisfy the following restriction:

$$\begin{aligned} 1 &= -\frac{\text{Cov}\left(dp_t, \sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j}\right)}{V(dp_t)} + \frac{\text{Cov}\left(dp_t, \sum_{j=1}^{\infty} \rho^{j-1} r_{t+j}\right)}{V(dp_t)} \\ &= -\beta_d + \beta_r \end{aligned}$$

where β_d and β_r are implicitly defined in the above.

Vector autoregressions (VAR)

- consider 1st order restricted VAR:

$$\begin{aligned} r_{t+1} &= a_r + b_r dp_t + \varepsilon^r_{t+1} \\ \Delta d_{t+1} &= a_d + b_d dp_t + \varepsilon^d_{t+1} \\ d_{t+1} - p_{t+1} &= a_{dp} + \phi dp_t + \varepsilon^{dp}_{t+1} \end{aligned}$$

- remember we log-linearized an identity to get:

$$r_{t+1} = \Delta d_{t+1} + \rho pd_{t+1} + \kappa_0 - pd_t.$$

with

$$\rho = \frac{e^{\overline{pd}}}{1 + e^{\overline{pd}}}$$

- this implies that there exists a deterministic relationship between these variables.

Cross-equation restrictions

- take expectations:

$$E_t[r_{t+1}] = E_t[\Delta d_{t+1}] + \rho E_t[pd_{t+1}] + \kappa_0 - pd_t.$$

- go back to the 1st order VAR:

$$\begin{aligned} E_t[r_{t+1}] &= a_r + b_r dp_t \\ E_t[\Delta d_{t+1}] &= a_d + b_d dp_t \\ E_t[d_{t+1} - p_{t+1}] &= a_{dp} + \phi dp_t \end{aligned}$$

- this implies that:

$$a_r + b_r dp_t = a_d + b_d dp_t - \rho(a_{dp} + \phi dp_t) + \kappa_0 - pd_t$$

Cross-equation restrictions

$$
\begin{aligned}
r_{t+1} &= a_r + b_r dp_t + \varepsilon^r_{t+1} \\
\Delta d_{t+1} &= a_d + b_d dp_t + \varepsilon^d_{t+1} \\
dp_{t+1} &= a_{dp} + \phi dp_t + \varepsilon^{dp}_{t+1}
\end{aligned}
$$

- $\Rightarrow$ the coefficients in these three equations must obey:

$$b_r = b_d + 1 - \phi\rho$$

  or equivalently that the following is true:

$$\frac{b_r}{1-\rho\phi} - \frac{b_d}{1-\rho\phi} = 1$$

  - the first term is the slope coefficient in the regression of the discount rate component on the *dp*-ratio
  - the second term is the slope coefficient in the regression of the cash flow component on the *dp*-ratio
  - we show this on the next slide

Slope coefficient background math

Consider two hypothetical regressions:

1. the cash flow component on the *dp*-ratio:

$$\sum_{j=1}^{\infty} \rho^{j-1} E_t \left[\Delta d_{t+j}\right] = \alpha_d + \beta_d dp_t + \varepsilon_t$$

   Substitute in for future dividend growth using the VAR specification (note error term equals zero always):

$$\sum_{j=1}^{\infty} \rho^{j-1} \left(a_d + b_d E_t \left[dp_{t+j-1}\right]\right) = c + \sum_{j=1}^{\infty} \rho^{j-1} b_d \phi^{j-1} dp_t = c + \frac{b_d}{1-\rho\phi} dp_t.$$

   Thus, $\beta_d = \frac{cov\left(dp_t, \sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j}\right)}{V[dp_t]} = \frac{b_d}{1-\rho\phi}$ (and c is a constant term).

2. the discount rate component on the *dp*-ratio:

$$\sum_{j=1}^{\infty} \rho^{j-1} E_t \left[r_{t+j}\right] = \alpha_r + \beta_r dp_t + \varepsilon_t$$

   Similar math as above yields $\beta_r = \frac{cov\left(dp_t, \sum_{j=1}^{\infty} \rho^{j-1} r_{t+j}\right)}{V[dp_t]} = \frac{b_r}{1-\rho\phi}$.

Cross-Equation Restrictions

Table: Cross-Equation Restrictions

	estimate	standard error	*implied*
b_r	0.105	[0.050]	0.114
b_d	0.007	[0.041]	-0.0017
ϕ	0.925	[0.056]	0.935

Notes: Annual Data. Sample 1927-2009. R_{t+1} is the real return on the CRSP-VW index. R_{t+1} denotes logs of the real return. R^f_{t+1} denotes the return on the real risk-free. ρ is .9650. The column calculates each coefficient based on the other two coefficients and the identity $b_r = 1 - \rho\phi + b_d$.

Variance Decomposition

- slope coefficients in predictability regressions represent fractions of variance due to discount rates and cash flows:

$$\frac{b_r}{1-\rho\phi} - \frac{b_d}{1-\rho\phi} = 1$$

  where we have assumed AR(1) for dividend yield with coefficient ϕ, b_r is slope coefficient in return regression, b_d is slope coefficient in dividend growth regression
- plugging in our estimates:

$$\frac{b_r}{1-\rho\phi} = \frac{0.105}{1-.9650\times.925} = .97$$

- discount rates account for 97 % of the variance of the log price/dividend ratio
  - Thus: the market valuation ratio moves around mostly because discount rates (expected returns) vary over time!

Cross-Equation Restrictions with Adjusted Dividend Yield

Table: Cross-Equation Restrictions

	estimate	standard error	*implied*
b_r	0.267	0.072	0.297
b_d	0.039	0.053	0.008
ϕ	0.768	0.768	0.800

Notes: Annual Data. Sample 1927-2008. R_{t+1} is the real return on the CRSP-VW index. R_{t+1} denotes logs of the real return. R^f_{t+1} denotes the return on the real risk-free. ρ is .965. The "implied" column calculates each coefficient based on the other two coefficients and the identity $b_r = 1 - \rho\phi + b_d$. This table uses the adjusted dividend yield $\widetilde{dp}$.

Variance Decomposition

- suppose we use the adjusted dividend yield instead
- go back to variance decomposition:

$$\frac{b_r}{1-\rho\phi} - \frac{b_d}{1-\rho\phi} = 1$$

  where we have assumed AR(1) for dividend yield with coefficient ϕ, b_r is slope coefficient in return regression, b_d is slope coefficient in dividend growth regression
- plugging in our estimates:

$$\frac{b_r}{1-\rho\phi} = \frac{0.267}{1-.965\times .768} = 1.0314$$

- discount rates account for 103 % of the variance of the log price/dividend ratio

The dog that did not bark

- if you believe $b_r = 0$, then you cannot believe $b_d = 0$

$$0 = b_d + 1 - \phi\rho$$

unless you think $\phi = \rho^{-1} > 1$

- so you need to explain the lack of evidence for dividend growth predictability, see Cochrane (2008)

The dog that did not bark

- Inspector Gregory:

  'Is there any other point to which you would wish to draw my attention?'

- Sherlock Holmes:

  'To the curious incident of the dog in the night-time.'

  'The dog did nothing in the night time.'

  'That was the curious incident.'

  (From "The Adventure of Silver Blaze" by Sir Arthur Conan Doyle.)

The Origins of Return Predictability

- different scenarios
  1. dividend/price ratio is persistent: some predictability
     - ★ $\rho = .9647$
     - ★ $\phi = .90$
     - ★ return regression slope coefficients

       $$b_r = 1 - \rho\phi = .139$$

       close to .097
  2. no persistence in dividend price ratio: too much predictability
     - ★ $\rho = .9647$
     - ★ $\phi = 0$
     - ★ return regression slope coefficients

       $$b_r = 1$$
  3. unit root in dividend price ratio: no predictability

     $$b_r \approx 0$$

Predictability of Returns

Return forecastability follows from the fact that dividends are not forecastable and that the dividend/price ratio is highly but not completely persistent.

Conclusion

- quite some evidence of stock return predictability
- other variables also predict stock returns (like the price/earnings ratio, risk-free rate, the yield spread etc.)
- evidence of predictability is much weaker in out-of-sample; see, Goyal and Welch (2008)

References

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay (1997). The Econometrics of Financial Markets. Princeton, NJ: Princeton University Press.

Campbell, J. Y. and R. J. Shiller (1989). The dividend-price ratio and expectations of future dividends and discount factors. The Review of Financial Studies 1(3), 195-228.

Cochrane, J. H. (2008). The dog that did not bark: A defense of return predictability. The Review of Financial Studies 21(4), 1533-1575.

Diebold, F. X. (2015). Forecasting. Department of Economics, University of Pennsylvania. http://www.ssc.upenn.edu/ fdiebold/Textbooks.html.

Lettau, M. and S. Van Nieuwerburgh (2008). Reconciling the return predictability evidence. The Review of Financial Studies 21(4), 1607-1652.

Stambaugh, R. F. (1999). Predictive regressions. Journal of Financial Economics 54(3), 375-421.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. The Review of Financial Studies 21(4), 1455-1508.

Appendix

Campbell and Shiller (1989) approximation

Let $p_t = \log(P_t)$ and $d_t = \log(D_t)$.

$$
\begin{aligned}
r_{t+1} &= \log(1 + R_{t+1}) = \log(\frac{P_{t+1} + D_{t+1}}{P_t}) = \log(P_{t+1} + D_{t+1}) - \log(P_t) \\
&= \log(P_{t+1} + D_{t+1}) - p_{t+1} + p_{t+1} - p_t \\
&= \log(\frac{P_{t+1} + D_{t+1}}{P_{t+1}}) + p_{t+1} - p_t \\
&= \log(1 + \exp(d_{t+1} - p_{t+1})) + p_{t+1} - p_t
\end{aligned}
$$

- Let $f(x) = \log(1 + \exp(x))$ be the function we need to approximate, i.e. $f(d_{t+1} - p_{t+1}) = \log(1 + \exp(d_{t+1} - p_{t+1}))$.
- We want to take a first-order Taylor series expansion around the unconditional mean $\bar{dp}$.

Campbell and Shiller (1989) approximation

The first-order Taylor series expansion around $\bar{dp}$ is

$$f(d_{t+1} - p_{t+1}) \approx \log(1 + \exp(\bar{dp})) + \frac{\exp(\bar{dp})}{1 + \exp(\bar{dp})}(d_{t+1} - p_{t+1} - \bar{dp})$$

Plug this expression in to the equation for r_{t+1} above

$$\begin{aligned} r_{t+1} &\approx \log(1 + \exp(\bar{dp})) + \frac{\exp(\bar{dp})}{1 + \exp(\bar{dp})}(d_{t+1} - p_{t+1} - \bar{dp}) + p_{t+1} - p_t \\ &= \log(1 + \exp(\bar{dp})) - \frac{\exp(\bar{dp})}{1 + \exp(\bar{dp})}\bar{dp} + \frac{1}{1 + \exp(\bar{dp})}p_{t+1} \\ &\quad + \frac{\exp(\bar{dp})}{1 + \exp(\bar{dp})}d_{t+1} - p_t \end{aligned}$$

This gives us the expression

$$r_{t+1} \approx \kappa + \rho p_{t+1} + (1 - \rho)d_{t+1} - p_t$$

where $\rho = \frac{1}{1+\exp(\bar{dp})}$ and $\kappa = \log(1 + \exp(\bar{dp})) - \frac{\exp(\bar{dp})}{1+\exp(\bar{dp})}\bar{dp}$.

Appendix: Forecasting

Forecasting

- suppose we want to forecast the value of Y_{t+1} based on a set of predictor variables $\mathbf{X}_t$

Example

If we want to use the last m values of Y_t, then
$\mathbf{X}_t = (Y_t, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-m+1})$

- let $Y^*_{t+1} = g(\mathbf{X}_t)$ denote any forecast function of $\mathbf{X}_t$ consisting of all the variables observed up until time t
- Which function $g(\mathbf{X}_t)$ of the observed data should we choose?

Loss functions

- to evaluate different forecast functions $g(\mathbf{X}_t)$, we need a systematic way to define what 'best' means.
- a loss function $L(Y_{t+1}, Y^*_{t+1})$ describes how we feel about missing our target Y_{t+1} if we choose our forecast to be $Y^*_{t+1} \; = \; g(\mathbf{X}_t)$.

  Examples:

  - squared error loss:

$$L(Y_{t+1}, Y^*_{t+1}) \; = \; (Y_{t+1} - Y^*_{t+1})^2$$

  - absolute deviation loss:

$$L(Y_{t+1}, Y^*_{t+1}) \; = \; |Y_{t+1} - Y^*_{t+1}|$$

Optimal forecasts

The **optimal forecast** is the function $Y^*_{t+1} = g(\mathbf{X}_t)$ that minimizes the expected loss

$$\underset{Y^*_{t+1}}{argmin}\ E[L(Y_{t+1}, Y^*_{t+1})|\mathbf{X}_t]$$

- minimizing expected loss is analogous to maximizing expected utility
- the field of **decision theory** studies optimal decision making under uncertainty.

MSE forecasts

Consider the squared error loss function:

$$L(Y_{t+1}, Y^*_{t+1}) \;=\; (Y_{t+1} - Y^*_{t+1})^2$$

The forecast that minimizes the mean squared error

$$\underset{Y^*_{t+1}}{argmin}\; E[(Y_{t+1} - Y^*_{t+1})^2 | \mathbf{X}_t]$$

is the **conditional expectation** of Y_{t+1} given by

$$Y^*_{t+1} \;=\; g(\mathbf{X}_t) \;=\; E[Y_{t+1} | \mathbf{X}_t]$$

- if our loss-function is SE, the **conditional expectation** is optimal

Forecasting

- mean squared error (MSE) is the default choice and is often implicitly assumed.
- There is a large literature in econometrics and statistics on optimal forecasting. Unfortunately, I am only scratching the surface.
- See the (free) book on forecasting: Diebold (2015)

Linear Forecasting

- **suppose we do not commit to a particular process for Y_t**
- suppose we want to forecast the value of Y_{t+1} based on a set of variables $\mathbf{X}_t$ but we only use linear forecasts $\mathbf{X}_t\alpha$

Definition

The **linear projection** of Y_{t+1} on $\mathbf{X}_t$ is defined by finding a vector $\boldsymbol{\alpha}$ such that the forecast error is uncorrelated with $\mathbf{X}_t$:

$$E[\mathbf{X}_t'(Y_{t+1} - \mathbf{X}_t\boldsymbol{\alpha})] = 0$$

Forecasting

The linear forecast $g(\mathbf{X}_t) = \mathbf{X}_t \boldsymbol{\alpha}$ that minimizes the mean squared error

$$\underset{\boldsymbol{\alpha}}{argmin} \;\; E[(Y_{t+1} - \mathbf{X}_t \boldsymbol{\alpha})^2 | \mathbf{X}_t]$$

is the linear projection of Y_t on $\mathbf{X}_t$.

- hence, the best linear forecast satisfies:

$$E[\mathbf{X}_t'(Y_{t+1} - \mathbf{X}_t \boldsymbol{\alpha})] = 0$$

- in other words, the linear projection coefficient satisfies:

$$E(\mathbf{X}_t' Y_{t+1}) = E(\mathbf{X}_t' \mathbf{X}_t) \boldsymbol{\alpha}$$

- hence, the projection coefficient is:

$$\boldsymbol{\alpha} = E(\mathbf{X}_t' \mathbf{X}_t)^{-1} E(\mathbf{X}_t' Y_{t+1})$$

Linear Projection and OLS

- consider the linear regression model:

$$y_{t+1} = \mathbf{x}_t \beta + u_{t+1}$$

- the sample sum of squared residuals:

$$\sum_{t=1}^{T} \left(y_{t+1} - \mathbf{x}_t \hat{\beta} \right)^2$$

- the value of $\hat{\beta}$ that minimizes the SSR is the OLS estimate:

$$\hat{\beta} = \left[\sum_{t=1}^{T} \mathbf{x}_t' \mathbf{x}_t \right]^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t' y_{t+1} \right]$$

Linear Projection and OLS

- OLS regression coefficients yield consistent estimates of the linear projection coefficients

- sound basis for OLS regressions when forecasting

Comparison with OLS

- recall our OLS estimator of $\boldsymbol{\beta}$:

$$\hat{\beta} = \left[\sum_{t=1}^{T} \mathbf{x}_t' \mathbf{x}_t\right]^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t y_{t+1}\right]$$

- hence, the OLS estimator uses *sample moments*, while the linear projection uses *population moments*

if $\{\mathbf{X}_t, Y_t\}$ are covariance-stationary, then the sample moments will converge to the population moments and the OLS estimator $\hat{\beta}$ will converge to $\boldsymbol{\alpha}$ as $T \rightarrow \infty$

- sound basis for forecasting under very mild conditions!!
- Note that this does not mean OLS is efficient (maximum likelihood is; GLS)