TRINITY COLLEGE DEPARTMENT OF ECONOMICS
WORKING PAPER 13-13

The Excess Returns Puzzle in Currency Markets:
Clues on Moving Forward

Josh R. Stillwagon

December 2013

Abstract

The paper is designed to review the empirical literature on the excess returns puzzle: the difficulty encountered by standard risk premium models in accounting for relative returns in the foreign exchange market. Of particular interest are the studies using survey data to decompose ex post excess returns into an expected component - the risk premium - and a forecast error. On the whole, these studies have found evidence of violations of the rational expectations hypothesis (non-white noise forecast errors), and a time-varying risk premium. This suggests the need for an alternative specification of forecasting. The literature has left open however the question of whether the traditional models can account for movements in the premium as measured by survey. Although the traditional models have not been tested against survey data, there is reason from the outset to believe EUT provides a deficient foundation for a model of the risk premium. Experimental evidence is discussed showing that the predictions of EUT are grossly inconsistent with the behavior of actual subjects towards risky gambles. Lastly, the paper discusses alternative models of risk preferences drawing from the experimental findings on prospect theory, and ways in which their testing can be improved through use of the I(2) cointegrated VAR model.

J.E.L. Codes: F31, D81, D84, G02, G15
Keywords: Excess returns puzzle, survey data, prospect theory, cointegration
THE EXCESS RETURNS PUZZLE IN CURRENCY MARKETS: CLUES ON MOVING FORWARD

by Josh Stillwagon\textsuperscript{a,b}

\textsuperscript{a}Department of Economics, Trinity College, Hartford, CT 06106
\textsuperscript{b}INET Center for Imperfect Knowledge Economics, University of Copenhagen
September 2013

Abstract

The paper is designed to review the empirical literature on the excess returns puzzle: the difficulty encountered by expected utility theory (EUT) and rational expectation hypothesis (REH) based risk premium models in accounting for relative returns in the foreign exchange market. Of particular interest are the studies using survey data to decompose ex post excess returns into an expected component - the risk premium - and a forecast error. On the whole, these studies have found evidence of violations of the rational expectations hypothesis (non-white noise forecast errors), and a time-varying risk premium. This suggests the need for an alternative specification of forecasting. The literature has left open however the question of whether the traditional models can account for movements in the premium as measured by survey. Although the traditional models have not been tested against survey data, there is reason from the outset to believe EUT provides a deficient foundation for a model of the risk premium. Experimental evidence is discussed showing that the predictions of EUT are grossly inconsistent with the behavior of actual subjects towards risky gambles. Lastly, the chapter discusses alternative models of risk preferences drawing from the experimental findings on prospect theory, and ways in which their testing can be improved through use of the I(2) Cointegrated VAR model.

\textbf{J.E.L. Code:} F31, D81, D84, G02, G12

\textbf{Keywords:} Excess returns puzzle, survey data, cointegration, imperfect knowledge economics (IKE) gap model

\textbf{Acknowledgements:} I would like to thank my advisor Michael D. Goldberg, and my dissertation committee members Mostafa Beshkar, Bruce Elmslie, Roman Frydman, and Le Wang for very helpful comments.
1 Introduction

Much of the theory that financial economists have developed over the past five decades to understand the behavior of relative, or excess, returns in asset markets is based on expected utility theory (EUT) and the rational expectations hypothesis (REH). It is uncontroversial to say that the empirical performance of these theories has been dismal.\(^1\) Most empirical studies make use of data on ex-post returns as a proxy for expected returns, invoking REH’s implication of white-noise forecast errors. This allows for only indirect tests of the models’ implications for ex-ante returns, which derive from the specification of risk preferences. The wide-spread rejection of these models raises a key question: does the failure arise from their specification of risk preferences and/or the assumption of REH?

In this chapter, I survey the empirical record for clues on this question and directions for further needed inquiry. I focus primarily on the research concerning currency markets, which arguably has made the most progress in uncovering the source of the failure; in part because of the availability of survey data on traders’ exchange rate forecasts. This has enabled researchers to decompose ex-post returns into an expected component, the risk premium, and an unexpected component, the forecast error. These studies find much evidence against REH. The forecast errors are not a white-noise process and appear to be correlated with prior information such as the forward rate. Some of these studies also report little evidence that a risk premium accounts for movements in expected returns. This has led some to the conclusion that risk plays little role and that market participants are grossly irrational, perpetually forgoing obvious profit opportunities as simplistic as merely betting against the forward rate.

I argue in this chapter that

1.) The survey data evidence, when taken as a whole, implies that risk does matter. Studies which conclude that risk doesn’t matter pool data across exchange rates and time horizons, and do not allow for structural change in the relationship. Other studies using un-pooled data and/or allowing for structural change find much evidence of a time-varying risk premium.

2.) The evidence against REH should not be interpreted as implying irrationality. Most of these studies do not allow for structural change in the correlations between fundamentals and the forecast error. The Contingent Expectations Hypothesis (CEH) of Frydman and Goldberg (2013) predicts that, given intermittent revisions in forecasting strategies, the relationship driving market outcomes in not likely to be time invariant. Testing for such parameter instability is an important aspect of determining the robustness of any econometric conclusion. If the correlations between fundamentals and the forecast error are changing over time, this belies the notion that simple profit-generating rules exist and that participants are forgoing obvious profit opportunities.

---

\(^1\)This literature is outlined briefly in the beginning of the chapter. For additional surveys see Lewis (1995) and Engel (1996).
3.) Although survey data has been used to test REH, very little research has been conducted with this data to examine the correspondence between the ex ante risk component of standard models and the expected returns of actual traders. The empirical record implies that allowing for structural change will be important here.

4.) Although standard models have not been tested against survey data, there is reason from the outset to believe they are deficient. These models rely solely on EUT to model preferences and decision making, the predictions of which have been found to be grossly inconsistent with subjects’ actual behavior towards risk (e.g. Kahneman and Tversky’s (1979, 1992) Prospect Theory). Progress has been made in developing alternative risk premium models, but much of the testing is based on ex post returns, which again confounds inference. An open question in the literature remains: which model can best account for the ex ante returns of market participants?

5.) One study, Frydman and Goldberg (2007), uses survey data to test the ex ante predictions of their Keynes-Imperfect Knowledge Economics (IKE) gap model which draws on the experimental findings of Prospect Theory. Extension of this work is warranted in a few directions however. The empirical work on the Keynes-IKE gap model could benefit from improved testing to better address non-stationarity and endogeneity, to explicitly test all of restrictions imposed on the data, and to better examine the driving and adjustment process of the relationship. It would also be of interest to conduct nested testing of alternative models, to provide direct individual comparisons of alternative theories, as well as to estimate the impact of multiple hypothesized influences on the premium simultaneously.

2 The Early Empirical Literature on Currency Returns

Much of the early empirical work on currency returns was aimed at testing the simplest model of ex ante returns, Uncovered Interest Parity (UIP), using ex post returns. This condition remains one of the core building blocks of the monetary approach to exchange rate determination, which continues to play an important role in the literature. UIP is derived under the assumptions of EUT and risk neutrality, perfect capital mobility, and homogenous expectations. The basic setup starts from a standard investment decision, how much of non-monetary wealth one should allocate between domestic and foreign assets in their portfolio.

2.1 Portfolio Balance under Risk Neutrality

We assume that there are two countries, A (referred to as domestic) and B (foreign), and two types of non-monetary assets, A and B bonds. Bonds of each
type are denominated in the currency of their respective countries. The riskless nominal returns to each bond from time $t$ to $t+1$ are denoted by $i^A_t$ and $i^B_t$ respectively. There are no barriers to short selling so individuals can issue A and B bonds (borrow domestic and foreign currency, respectively). Using log approximations the ex post nominal return on a pure long position in foreign exchange held one period, $r^L_{t+1}$:

$$r^L_{t+1} = s_{t+1} - s_t + i^B_t - i^A_t \tag{1}$$

where $s_t$ denotes the log level of the time-$t$ spot rate.\(^2\) The ex post nominal return on a pure short position in foreign exchange $r^S_{t+1}$ is simply the negative of the long position. Equilibrium in the foreign exchange market is defined by a balance between the demand for and supply of foreign exchange. In deriving the condition for momentary equilibrium, it is assumed that the equilibriums in domestic and foreign money markets occur independently of the spot-rate process and the level of wealth, and that domestic and foreign currency are held only by residents of those countries.\(^3\) The first assumption implies that equilibrium in the foreign exchange market can be modeled in terms of domestic and foreign individuals’ decisions as to how to divide their non-monetary wealth between domestic and foreign bonds. The second assumption ensures that whenever an individual alters their portfolio (relative holdings of domestic and foreign bonds) they will buy or sell foreign currency.

$$r^i_{t+j+1} = s^i_{t+j+1} - s_t + i^B_t - i^A_t \tag{2}$$

The assumption of risk neutrality yields a particularly simple portfolio-decision rule for all individuals. They are concerned only with the expected first moments of returns. Thus individuals’ decision rule is simply to hold only the asset with the highest expected return, irrespective of the other moments of the returns distribution. If $r^i_{t+j+1} > 0$ ($< 0$), an individual would want to issue A bonds (B bonds) in order to use the proceeds to increase their holding of B (A) bonds without limit. That is:

This incipient capital flow from any expected excess return, given perfect capital mobility, would drive down the nominal yield (interest rates) on that asset, and bid up its currency value until the expected excess return disappeared. In equilibrium then, UIP implies that expected returns equalize across all assets ($r^i_{t+1} = 0$), in which case individuals have no incentive to purchase or sell.

\(^2\)To take a pure long position in foreign exchange for one period, which requires no money down at time $t$, an individual would borrow, say, one unit of domestic currency at time $t$ at a rate of $i^A_t$, then immediately sell this at $1/s_t$ in the spot market lending $1/s_t$ units of foreign currency at $i^B_t$. At $t + 1$ the individual will sell $(1 + i^B_t)s_{t}$ units of foreign currency at $s_{t+1}$. The return using log approximations yields the above equation.

\(^3\)Branson and Henderson (1985) include these assumptions in their "basic asset market specification".
foreign exchange in order to alter their portfolio. Another way to state this is that in equilibrium any interest rate differential should be exactly offset by one-for-one expected depreciation.

2.2 Tests of UIP assuming REH

In order to test this proposition \((\hat{r}_{jt+1} = 0)\), economists in practice relied on the data for ex post returns \(r_{t+1}\), since \(\hat{r}_{jt+1}\) is generally unobservable. This requires one to specify the connection between ex ante and ex post returns, or the process for the forecast error, as you will recall from equation (1) \(r_t = \hat{r}_{jt+1} + \varepsilon_{t+1}\).

It was natural for economists to believe that market participants behaved in mostly reasonable, rational ways. Consequently, they assumed the forecast error was a purely white-noise error, as opposed to assuming some arbitrary, yet specific, process of systematic misforecasting. As a result, these tests are a joint hypothesis of UIP \((\hat{r}_{jt+1} = 0)\) and REH \((\varepsilon_{t+1} \sim i.i.d. N(0, \Omega))\), yielding the prediction that \(r_{t+1}\) should be white noise. Obstfeld (1985) examines the statistical properties of \(r_{t+1}\) and finds significant evidence that it is serially correlated. This suggests that, if stable, returns are actually predictable.

An alternative means to indirectly test the implications of UIP with REH is to examine whether the interest rate differential (forward discount) is associated with one-for-one exchange rate depreciation. This is one way to test whether returns are in fact white-noise or whether they are correlated with time \(t\) information (here the forward discount).

\[
\Delta s_{t+k} = \alpha + \beta f^k_t + \varepsilon_{t+k} \tag{3}
\]

where \(\Delta s_{t+k}\) is the percentage depreciation of the exchange rate over \(k\) periods (the change in the log of the spot exchange rate), and \(f^k_t\) is the current \(k\)-period forward premium (the log of the forward rate minus the log of the spot rate). This equation will be referred to as the Bilson-Fama (BF) regression. The null hypothesis of unbiasedness is that \(\alpha = 0\) and \(\beta = 1\). A finding that the forward rate is an unbiased predictor of the future spot rate implies both REH (white-noise forecast errors) and risk neutrality (or less strongly perfect substitutability between the two countries’ bonds). As a result this is sometimes referred to as a test of the risk-neutral, efficient markets hypothesis.

Froot (1990) surveys 75 studies testing this hypothesis, all of which reject forward rate unbiasedness. Further, the estimate of \(\beta\) is consistently less than one, in fact most are negative indicating that the exchange rate actually moves in the opposite direction from that predicted by the forward rate and UIP, with an average estimate of -.88. A few are positive but not one has a point estimate equal to or greater than one. Although the estimates of \(\alpha\) themselves are often treated as of secondary importance, they are in several instances statistically significant in both Bilson (1981) and Fama (1984). This rejection of UIP and REH, if stable, suggests that one could receive an excess return on average simply by betting against the forward rate (investing in the higher interest rate country). This result became known as the forward discount anomaly.
One possible explanation for the bias was that it could arise due to paradox noted by Siegel (1972). If \( F_t \) is an unbiased estimate of \( S_{t+1} \), it is impossible for \( 1/F_t \) to be an unbiased estimate of \( 1/S_{t+1} \) since by Jensen’s inequality \( E(1/S_{t+1}) \neq 1/E(S_{t+1}) \) due to the convexity of the reciprocal form. Empirical work by McCulloch (1975) and others however has demonstrated that the Jensen’s inequality term (JIT) is far too small to account for the observed bias. Most empirical work following has worked in natural logarithms to largely circumvent this issue.

2.3 Ambiguity of Interpretation

While researchers typically agree on the presence of the forward rate bias, although as discussed momentarily there is reason to believe this bias is not uniform over time, there is much contention over its source. Some assume that individuals are risk neutral (or rather all risk is diversifiable), and in turn the bias would be associated with expectational errors. Bilson (1981) referred to this as excessive speculation since, under this interpretation, individuals would do better to put a larger weight on the current spot rate than all other information in their forecasting strategy.

Others meanwhile assume individuals would not make systematic errors over the sample, and in turn view the bias as evidence of a time-varying risk premium. There is of course no way to differentiate between these two explanations without further information about the expected return, since these estimates of a forward discount have been a joint test of both the ex ante component (here UIP) and the assumption of white-noise forecast errors (REH).

There are statistical issues however which would lead one to question whether it is even appropriate to interpret these results as implying predictable excess returns. Typically the fit of these models is very poor. In Fama (1984) for example they are typically below 0.1 and never exceed 0.2 for any of the nine exchange rates examined. Another major issue is that the results seem to be sample dependent and subject to structural change. Hansen and Hodrick (1983) posited that the process driving flexible exchange rates may not have been well understand by market participants in the early portion of the samples in the 1970’s, given their recent implementation following the end of the Bretton Woods system in 1973. Both Fama (1984) and Lewis (1988) divide their samples into equal length sub-samples (two in the case of the former and three in the case of the latter) arriving at rather widely differing estimates in each case, though largely negative in Fama’s study. Meanwhile, Lewis in her later sub-samples finds evidence of positive estimates for \( \beta \).

Bekaert and Hodrick (1993) also split their sample for the BF regression and find negative estimates for \( \beta \) in the first half of their sample, and largely positive estimates in the second half of their sample. The difference appears to be statistically significant based on tests for one predetermined break point, though the break was chosen somewhat arbitrarily as the end of Bilson’s original sample. Goldberg and Frydman (1996) find parameter instability for the traditional monetary models relating exchange rates and fundamentals, using tests which
do not pre-specify the number or timing of regimes shifts (namely piece-wise linear relationships).

Frydman and Goldberg (2007) also find evidence of parameter instability in the BF regression suggesting that the coincidental negative correlation between the forward rate and a future depreciation disappears in alternative sub-samples, and in turn there is no reason to expect the profitability of betting against the forward rate to hold in the indefinite future, as it has often been interpreted. All of this evidence strongly demonstrates the importance of checking and addressing structural change to obtain meaningful inference. The forward discount anomaly and the interpretation that it implies gross, timeless irrationality on the part of market participants appear to be merely artifacts of failing to acknowledge structural change. It also suggests a need to depart from the assumption of REH and its presumption of a time-invariant conditional probability distribution.

There is also the issue of non-stationarity, which garnered much more attention after the Engel and Granger (1987) results on spurious regression and cointegration. Non-stationarity of the variables would lead to inconsistent estimates, and thus needs to be sufficiently addressed. Bilson (1981) finds significant auto-correlation in three of his nine samples based on the Durbin Watson statistics.

The persistence of the exchange rate however is actually a graver issue than is commonly recognized, or addressed by the Engel Granger error-correction model. Engel and Hamilton (1990) reject a simple random walk model of the exchange rate in favor of a broken trends model (which includes a time-varying trend which alternates between positive and negative values according to a Markov process), suggesting that the changes in the exchange rate are actually persistent. Growing evidence demonstrates that the traditional univariate tests for unit roots (such as the augmented Dickey-Fuller test) have low power to detect an I(2) component, or stochastic trend, when the noise-to-signal ratio is high, that is the variance of the I(1) component is much larger than that of the I(2) component. This large noise-to-signal ratio is often found to be the case (Juselius et al. 2012). This issue then likely plagues other tests on the persistence, or auto-correlation, of highly volatile series, such as the exchange rate or price levels. Alternative multivariate tests for I(2) behavior are advocated as a more powerful and appropriate alternative to investigate and detect for time-varying trends.

Others have found non-linearities in terms of threshold effects, wherein the exchange rate only tends to mean revert to its "fundamental level" beyond some threshold (Taylor, Peel and Sarno 2001). These types of non-stationarities imply that the linear BF results are potentially misleading. Baille and Bollerslev (2000) also demonstrate that the anomalous results from the BF regression could be due simply to a small sample bias. They show in simulation that even if UIP is imposed (a slope coefficient of one in the BF regression) that if the volatility is very persistent, the estimate will converge to the true value only very slowly.

---

4The broken-trends model is a special case of the more general I(2) model.
Further, they show that in sample sizes typical of the empirical literature, one can arrive at estimates similar to those often reported. This provides further motivation for the use of survey data on expected returns in that it circumvents some of the added volatility in ex post returns, deriving from the additional noise of the forecast error.

3 The Excess Returns Puzzle

The initial rejections of UIP with REH (i.e. forward rate unbiasedness or the risk-neutral efficient markets hypothesis) were not immediately devastating to the basic tenets of the field. Of course, few economists believed individuals were in fact risk neutral, and this in turn seemed like a very natural assumption to modify. There has been an enormous amount of research attempting to account for the reported predictability of returns with an REH risk premium model. To see how a risk premium could explain the results of the BF regression, consider the "true" model where in equilibrium, any expected excess return is attributable to a risk premium:

$$s_{i,t+1} - s_t = i_t^B - i_t^A - r_p = 0$$

where \(r_p\) is the risk premium, and now it is risk adjusted returns which equalize. If the foreign asset is viewed as riskier, there will be an excess return on foreign exchange (\(r_p > 0\)). To see how an omitted risk premium could bias the BF results, if we assume REH the BF regression instead estimates:

$$s_{i,t+1} - s_t = \hat{\beta}(f_{p_i}^k) + \varepsilon_{t+1}$$

$$\hat{\beta} = \beta + \frac{cov(f_{p_i}^k, r_p)}{var(f_{p_i}^k)} = 1 - \frac{cov(f_{p_i}^k, r_p)}{var(f_{p_i}^k)}$$

where \(f_{d_i}^k\), the forward discount, is equivalent to the interest rate differential \(i_t^A - i_t^B\). \(\hat{\beta}\) represents the regression estimate of \(\beta\), and \(\frac{cov(f_{p_i}^k, r_p)}{var(f_{p_i}^k)}\) represents the omitted variable bias. Given the assumption of white noise forecast errors, \(cov(f_{d_i}^k, \varepsilon_{t+1}) = 0\), the bias would be the covariance between the included regressor, here \(f_{p_i}^k\), and the omitted variable times its coefficient in the true model (here \(-r_p\)), divided by the variance of the included regressor. If the forecast error is white noise and individuals are risk neutral, the "true" beta would be one. It can be seen then that a covariance between the risk premium and forward discount would bias the estimate downward.

Recognition of this potential for a risk premium to account for the observed bias in the forward rate has led to extensive research attempting to answer the

\(^5\)This follows from covered interest parity, otherwise traders could arbitrage the difference and acquire riskless profits. This condition is widely observed to hold, and many banks set their forward contracts to be consistent with prevailing interest rates.
question "which risk premium model can account for the ex post bias of the forward rate?"

3.1 The Traditional Capital Asset Pricing Model

The earliest attempts to explain the forward discount anomaly with a risk premium adapted the portfolio balance approach or capital asset pricing model (CAPM) to the international context. As with the derivation of UIP, this is a two-country, two period setup where individuals allocate their non-monetary wealth between country A bonds, and country B bonds, both of which have random real returns, so as to maximize their next period’s expected utility.

If the probability distribution of real returns is normal, this implies that the distribution can be entirely described by its mean and variance, and in turn investors can compare portfolios on the basis of just their first two moments. Risk-averse preferences then are assumed to depend positively on the conditional mean of next period’s wealth, \( \hat{\mu}_{it+1}^w \) and negatively on its conditional variance, \( \hat{\sigma}_{it+1}^w \) both of which are implied by the economist’s representation of an individual’s forecasting behavior. The utility function then

\[
U = U(\hat{\mu}_{it+1}^w, \hat{\sigma}_{it+1}^w), \quad \frac{\partial U}{\partial \hat{\mu}_{it+1}^w} > 0, \quad \text{and} \quad \frac{\partial U}{\partial \hat{\sigma}_{it+1}^w} < 0 \tag{7}
\]

where the superscript i, denoting that the utility function and conditional moments above are for an individual, has been omitted to reduce the notation.

The next period’s expected real wealth is

\[
\hat{W}_{it+1}^i = W_{it+1}^i \left[ \hat{\mu}_{it+1}^w + (1 + i_t^A p_t^A) \right]
\]

where \( j = A, B \) and the remainder are defined as before. Utility maximization implies a portfolio share \( a_t \) corresponding to:

\[
a_t = \frac{\hat{\mu}_{it+1}^w}{\rho \hat{\sigma}_{it+1}^w} + d \quad \text{where} \quad a_t \in [0, 1] \tag{8}
\]

where \( \hat{\mu}_{it+1}^w \) is the conditional mean return and \( \hat{\sigma}_{it+1}^w \) denotes the conditional variance of the change in the exchange rate \( \Delta s_{t+1} = s_{t+1} - s_t \), \( \rho \) denotes the coefficient of relative risk aversion implied by the utility function, and \( d \) denotes the minimum-variance portfolio share of foreign bonds.

The expected utility maximizing portfolio share expressed above assumes that the rates of domestic and foreign inflation are deterministic, implying that

---

\( ^6 \)See for example Kouri (1976) and Dornbusch (1983) for seminal developments of the model.

\( ^7 \)Although the two-period setup is more restrictive than the intertemporal optimizing framework, Giovannini and Jorion (1989) show that it is consistent with intertemporal optimization if the elasticity of intertemporal substitution equals one.

\( ^8 \)See appendix for derivation.
\[ \hat{\nu}_{t+1}^{\Delta s} = \hat{\nu}_{t+1}^{\Delta s} \] We also assume that individuals use only their home country’s prices to deflate the nominal values. These two assumptions imply that the minimum-variance portfolio for all individuals contains bonds issued only by their respective countries. This minimum variance portfolio implies \( d = 0 \) for domestic residents, and \( d = 1 \) for foreign residents.\(^9\)

By definition then, a domestic individual holds an open position in foreign exchange when \( a_t \neq 0 \), while a foreign individual holds an open position when \( a_t \neq 1 \). The solution from the previous equation implies that both domestic and foreign wealth holders will want to hold a long position in foreign exchange of size \( (a_d)W \) whenever \( \hat{r}_{t+1} > 0 \), and a short position of size \( -(a_d)W \) whenever \( \hat{r}_{t+1} < 0 \). Further, the size of one’s open position is increasing in the forecasted return, and decreasing with the degree of risk aversion and the perceived risk involved, as measured by \( \hat{\nu}_{t+1}^{\Delta s} \).

The literature often assumes that market participants have homogenous forecasts and preferences. Given these assumptions, the momentary equilibrium condition (aggregate bond demands must equal aggregate bond supplies), and the expression for the optimal portfolio share, momentary equilibrium in the foreign exchange market can be expressed as:

\[ \hat{r}_{t+1} = r_p + \rho \hat{\nu}_{t+1}^{\Delta s} IFP_t \] (9)

where \( \hat{r}_{t+1} \) and \( \hat{\nu}_{t+1}^{\Delta s} \) are implied by the economist’s specification of the common forecasting strategy, \( IFP_t = \frac{B_t^A - A_t^B}{W_t} \) where \( B_t^A \) and \( A_t^B \) represent the total value of each country’s bonds held by non-residents entering time \( t \), and \( W_t = \sum W_t^i \) is the total non-monetary wealth in both countries. The \( IFP_t \) then is the international financial position of country A vis-a-vis country B, expressed as a proportion of the total market’s non-monetary wealth. If \( IFP_t > 0 \), country A is a net creditor to country B. This term informs the size of the net long position which must be held on foreign exchange (short if \( IFP_t < 0 \)).

The above equation, referred to as risk-adjusted uncovered interest parity (RAUIP), states that equilibrium occurs when the aggregate, risk-adjusted expected returns equalize, that is \( \hat{s}_{t+1} - s_t + i_t^B = i_t^A + \rho \hat{\nu}_{t+1}^{\Delta s} IFP_t \). As can be seen, foreign exchange would require an expected excess return in equilibrium when the foreign country is a net debtor (\( IFP_t > 0 \)), and the market must hold a net long position. This equilibrium risk premium depends positively on the degree of risk aversion, \( \rho \), their assessment of the riskiness of holding open positions, \( \hat{\nu}_{t+1}^{\Delta s} \), and the size of the net long positions which must be held, \( IFP_t \).

\(^9\)Lewis (1995) and Engel (1996) also use these assumptions. The assumption of deterministic inflation rates is typically justified by their much lower variance relative to the exchange rate, Krugman (1981).
3.2 Gross Empirical Inconsistencies of the CAPM

Empirical work testing the CAPM has typically involved estimating versions of the above equation for RAUIP with ex post data (assuming REH), and evaluating it based on two criteria. The first is to test the implied restriction of mean variance optimization against the more general model where the coefficient on the $IFP_t$ term is an unrestricted, time-varying parameter. A second important criteria is to examine the estimates of $\rho$. Theory implies that it should be positive and statistically significant, but it also needs to fall within what is generally regarded as a reasonably low range. Mehra and Prescott (1985) for example did not consider estimates higher than 10, while others argue a reasonable estimate would be around 2 (Krugman 1981). This corresponds to an individual being indifferent between a 4% loss, and a gamble equally likely to produce a gain or loss of 20%.

The various studies differ primarily in the way they model $\hat{\sigma}_{t+1}^2$. The earliest studies assumed the variance was a constant, while later studies allow the variance to vary over time according to options prices, fundamental variables, or an auto-regressive conditional heteroskedasticity (ARCH) process. Engel (1996) summarizes that the performance of the international CAPM models has largely been disappointing.

Lewis (1988a) estimates the portfolio balance model assuming a constant variance. She uses monthly data on government bond holdings (for the U.S., U.K., West Germany, Canada, and Japan), and Eurocurrency rates from January 1975 to December 1981. Lewis finds that relative asset supplies have little explanatory power for returns. The portfolio balance approach implies that an increase in a country's bonds should lead to a higher return, while two of the four estimates are negative.

Engel and Rodriguez (1989) and Giovanni and Jorion (1989) model the variance using an ARCH or GARCH model, which assumes the variance of the current error term is a function of the size of the previous error term, and alternatively allowing it to vary according to other economic data, including interest rates. Engel and Rodriguez use monthly data on real returns and government bonds for the U.S., France, Germany, Italy, Japan and the U.K. from April 1973 to December 1984. Giovanni and Jorion use weekly data on nominal returns and government bonds of the U.S., Germany, the U.K., and Switzerland, and also the U.S. stock market, from July 1974 to December 1986. Both studies find that the estimates of $\rho$ (estimated for multiple country's bonds) are either insignificantly different from zero, or negative (whereas it should be positive in the model), and the restrictions of CAPM are all strongly restricted.

Lyons (1988) allows the variance to vary, but measures the conditional variance by backing it out of options prices. Estimation is conducted for the mark, yen and pound rates with the US dollar, using monthly data on returns and asset values from April 1983 to December 1985. Again the estimates of $\phi$ are not significant, and in fact the point estimates are negative. Thomas and Wickens (1993) include data on equities as well as bonds to represent asset supplies for
Japan, Germany, the U.S. and the U.K. from June 1976 to December 1987. The estimates of $\phi$ are actually statistically significant but negative, and the CAPM restrictions are rejected.

One possibility is that the empirical applications of the CAPM have simply failed to input the relevant conditional volatility. An alternative to the GARCH models has been further developed and more commonly employed in recent years, referred to as realized volatility. The realized measures use intra-period observations to construct a measure of the variance, rather than assuming it is a function of the previous observations of the error term as in the ARCH models. For example to construct a daily measure, intra-day tick data would be employed, or to construct a monthly measure the data on daily returns would be used to calculate the realized monthly series.\(^{10}\) This unconditional calculation has been shown to be an unbiased and highly efficient estimator of volatility. Employing a realized measure to evaluate the CAPM in foreign exchange is a potentially fruitful avenue for future research extensions.

There is another aspect in which the international CAPM is more fundamentally inconsistent with the behavior of returns. The excess return, both ex post as well as ex ante measures based on regression estimates, undergo frequent sign reversals. Lewis (19995) for example uses the BF regression and Mark and Wu (1998) use a bivariate vector auto-regression to estimate the premium. In both cases, the series demonstrate quite frequent sign reversals. This can only occur in the CAPM when countries alter from net debtor to net creditor, which occurs very seldomly.

### 3.3 The Consumption Capital Asset Pricing Model

A popular avenue attempting to explain the forward discount anomaly via an REH risk premium has focused on the connection between returns and consumption, referred to as the consumption CAPM (or CCAPM).\(^{11}\) The important feature of preferences in the model is that the marginal rate of substitution for consumption varies inversely with consumption growth; that is when consumption is high, it’s marginal utility is low. As a result, an asset which payoffs in bad states (low consumption) would be valued more heavily all else equal, while an asset with pro-cyclical payoffs (coinciding with good states where consumption is high) would then require a premium in equilibrium. In this model, risk is solely connected to this covariance between pays off and consumption. If the covariance is zero (or viewed to be zero), no premium would exist in equilibrium, as idiosyncratic risk (variance) is not priced in the model.

In applied asset pricing research, per capita consumption is often used. This can be justified by assuming the utility function in the Lucas two-period model is homothetic and that the relative price of home and foreign goods (the real

---

\(^{10}\) See French, Schwert and Stambaugh 1987 for monthly measures, and Andersen, Bollerslev, Diebold and Labys 2003, and Berndorff-Nielsen and Shephard 2002 for daily measures.

\(^{11}\) The discussion in this section follows Engel (1996) and Mark (2001). See therein for further details.
exchange rate) is constant. While the latter assumption in particular is false, it allows one to derive more easily the basic conclusions of the model concerning a connection between the premium and the covariance between returns and consumption growth. The use of this assumption in empirical work however does appear dubious. The more general case will be discussed in subsequent sections. This assumption enables one to write the representative agent’s marginal rate of substitution between $t$ and $t+1$ as:

$$
\mu_{t+1} = \frac{\beta U'(C_{t+1})}{U'(C_{t})}
$$

(10)

where $U'(C_t)$ is the marginal utility of consumption evaluated at equilibrium, and $\beta$ is the subjective discount factor. A speculative position in the forward market requires no payment at time $t$. If the agent is behaving optimally, the expected marginal utility of further speculation will be zero, that is $E_t[U'(C_{t+1})(F_t - S_{t+1})/P_{t+1}]$. This Euler equation, the first-order condition, is typically rewritten by multiplying both sides by $\beta$ and dividing by $U'(C_{t+1})$.

$$
E_t[\mu_{t+1} \frac{(F_t - S_{t+1})}{P_{t+1}}] = 0
$$

(11)

The motivation for this rearrangement is in part said to be the need for an expression in terms of stationary random variables, so the conditional and unconditional variances and covariances between them exist. It is worth noting however that the observables variables in the expression however (prices, future exchange rates, and the forward rate) are often found to be non-stationary using multivariate tests (Juselius 2006).

We can then exploit the property of the covariance decomposition

$$
\text{Cov}_t(X_{t+1}, Y_{t+1}) = E_t(X_{t+1}, Y_{t+1}) - E_t(X_{t+1})E_t(Y_{t+1})
$$

(12)

If we substitute $X_{t+1} = \mu_{t+1}$ and $Y_{t+1} = \frac{(F_t - S_{t+1})}{P_{t+1}}$, we can see from equation (16) that the first term on the right-hand side of the covariance decomposition will equal zero and we can rearrange to obtain:

$$
E_t[\{F_t - S_{t+1}\}/P_{t+1}] = \frac{-\text{Cov}_t[\{F_t - S_{t+1}\}/P_{t+1}, \mu_{t+1}]}{E_t(\mu_{t+1})}
$$

(13)

This condition states that the expected forward payoff is inversely proportional to the conditional covariance between the payoff and the marginal rate of substitution (the latter of which varies inversely with consumption growth).
The factor of proportionality is $-1/E_t(\mu_{t+1})$ which is the expected gross real interest rate multiplied by $-1$.

Let’s begin by focusing on the interpretation under the case where $E_t[(F_t - S_{t+1})/P_{t+1}] < 0$. This implies two things. Firstly that individuals expect to make an excess return by holding a long position in foreign currency, (borrowing in the forward market and selling in the spot market at $t+1$, or borrowing home currency and lending uncovered in the foreign currency). Thus since foreign exchange is garnering a premium, it in equilibrium must be the riskier asset. Secondly, we can see that the covariance on the right-hand side is positive. The high payoffs occur when $(F_t - S_{t+1})/P_{t+1} < 0$ so these states are associated with low values of $\mu_{t+1}$, which occurs when consumption is high. Risk-averse investors however would prefer an asset that tends to pay off when consumption is low (since the marginal utility is higher). Since the payoff is procyclical, investors require a premium to hold foreign exchange in this case.

Since the foreign asset is riskier, the home asset must be safer. Given $E_t[(F_t - S_{t+1})/P_{t+1}] < 0$, you expect to experience a loss by holding a long position in the domestic asset. This may seem in conflict with logic, holding an asset one expects to lose money on, however given it tends to pay off when consumption is low ($\mu_{t+1}$ is high), the marginal utility of the payoff in terms of consumption is greater. The expected loss then can be viewed as a consumption insurance premium.\textsuperscript{12}

### 3.4 Gross Empirical Inconsistencies of the CCAPM

Mark (1985) provides the first study to estimate an equation similar to (12). In it he assumes constant relative risk aversion. Mark estimates the Euler equation using Hansen and Singleton’s (1982) Generalized Method of Moments (GMM) technique. GMM allows one to test the implications of the model without requiring as much information. Typically the focus is on the first order conditions, for example the Euler equation, which holds even if we do not know the process driving the shocks. The objective is to estimate the coefficient of relative risk aversion $\rho$ to, as in the CAPM, see if its estimate is positive, significant, and reasonable. The empirical tests on the Euler equation also evaluate the given orthogonality conditions. The Euler equation provides a necessary condition, so while failing to reject it may not imply that the model is consistent with the data in all respects, a rejection is sufficient to determine that the theory needs modification.

The model in Mark (1985) is estimated jointly for the Canadian dollar, mark, guilder and pound exchange rates with the U.S. dollar. Mark uses monthly data from March 1973 to July 1983, and aggregate per capita consumption (adjusted by monthly population estimates) measured as both non-durables, and non-durables and services. He uses the consumption deflator as the measure of prices.

\textsuperscript{12}If individuals are risk neutral, the marginal rate of substitution is constant, and since the covariance with a constant is zero, the Euler equation reduces to $E_t(F_t/P_{t+1}) = E_t(S_{t+1}/P_{t+1})$.\textsuperscript{12}
To condense the problem, let $r_{t+1}$ be the $4 \times 1$ vector of forward foreign exchange returns for the four exchange rates examined

$$r'_{t+1} = [(F_{1t} - S_{1t+1})/(S_{1t}), (F_{2t} - S_{2t+1})/(S_{2t}), (F_{3t} - S_{3t+1})/(S_{3t}), (F_{4t} - S_{4t+1})/(S_{4t})]$$  

(14)

and let $M_{t+1}$ be the $4 \times 1$ vector $M_{t+1} = \mu_{t+1} r_{t+1}$ where $\mu_{t+1}$ is the intertemporal marginal rate of substitution for the home representative investor under CRRA utility, $U(C) = C^{1-\rho}/(1-\rho)$, which would be $\beta(C_{t+1}/C_t)^{1-\rho}(C_{xt}/C_{xt+1})$ where the subscript $x$ denotes home country goods. With the Euler equation we get

$$E[M_{t+1}|I_t] = 0$$  

(15)

We can divide both sides by $\beta$ so only $\rho$ needs to be estimated, and the above equation states that $M_{t+1}$ should be uncorrelated with any time $t$ information (that is there are no risk-adjusted excess returns, risk-adjusted as defined by the CCAPM). In order to test this, let $z_t$ be a $k$-dimensional vector of available time $t$ "instrumental variables". This allows one to construct a system of $3 \times k$ equations to estimate and test

$$E[M_{t+1} \otimes z_t] = 0$$  

(16)

where $\otimes$ is the Kronecker product such that $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, and $B$ is any $n \times k$ matrix, then $A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{pmatrix}$

The question then becomes which variables to use as instruments. Mark (2001) notes that it is not a good idea to include too many variables, as the estimation will lose feasibility and the small sample properties will be harmed. A natural selection is to use the forward discount, due to its direct relevance to the puzzle at hand. One needn’t use all orthogonality conditions, so to further reduce the problem let for each currency $i$

$$z_{it} = \begin{pmatrix} 1 \\ (F_{it} - S_{it})/S_{it} \end{pmatrix}$$  

(17)

Mark then estimates the system of equations. His findings are emblematic of subsequent studies. The estimates for the coefficient of relative risk aversion $\phi$ are far larger than what is typically viewed as plausible, in cases exceeding 40. The restrictions are also rejected, that is the residuals (risk-adjusted excess returns) are not orthogonal to the information set, but rather forecastable from the forward discount.
Several studies update and expand upon Mark (1985). Hodrick (1989b) uses quarterly data on seven U.S. dollar and British pound exchange rates from 1973:3 to 1987:4 with the mark, pound, Canadian dollar, Belgian franc, French franc, guilder, and Swiss franc. For the U.S. exchange rates the consumption measure is for non-durables and services, while for the U.K. exchange rates it is non-durables alone. He does not reject the over-identifying restrictions in either case, but the estimate of \( \rho \) for the US data is 60.9, though it is a quite reasonable 2.15 with the U.K. data.

Modjtahedi (1991) conducts a study similar to Mark (1985), but testing the entire term structure of available forward rates (one-, three-, six-, and twelve-month maturities). The data is monthly from July 1973 to July 1988 for the pound, Canadian dollar, mark and yen relative to the dollar. Following Mark again, Modjtahedi uses both consumption of non-durables, and nondurables and services. The model is similarly rejected, and the lagged forward discounts appear to have explanatory power.

Kaminsky and Peruga (1990) use a sample for the mark, yen and pound with the U.S. dollar on monthly data from April 1975 to June 1985, with U.S. expenditure on non-durables plus services, and the consumption deflator. Under rational expectations the forward rate bias can be written as:

\[
s_{jt+1} - f_t = \frac{-1}{2} Var_t(s_{jt+1}) + Cov_t(s_{jt+1}, p_{t+1}) + \rho Cov_t(s_{jt+1}, c_{t+1}) + \varepsilon_{jt+1} \tag{18}\]

for each currency \( j \). They then stack the currencies into a vector, along with consumption and prices, which Kaminsky and Peruga write as:

\[
z_{t+1} = \beta_0 + \beta_1(L)z_t + D_0vec(H_{t+1}) + \nu_{t+1} \tag{19}\]

where \( z_{t+1} = [c_{t+1} - c_t, \ p_{t+1} - p_t, \ (s_{t+1} - f_t)^{BP}, \ (s_{t+1} - f_t)^{GM}, \ (s_{t+1} - f_t)^{JY}] \)

\( H_{t+1} \) is the covariance matrix of the vector of errors \( \nu_{t+1} \), which is modeled as a vector version of Bollerslev’s (1986) GARCH model. They then estimate the model with maximum likelihood according to the implied restrictions and find an estimate of \( \rho = 372.4 \), though with a standard error of 274.2 (note the likelihood ratio test rejects \( \rho = 0 \), contrary to the Wald test results). While the restrictions of a coefficient of -1/2 and 1 on the variance and covariance between the exchange rate and prices respectively is not rejected, the estimated residuals are not orthogonal to all time t information, rather they are forecastable using the forward discount.

Mark (2001) provides some intuition for the large estimates of \( \rho \) by reformulating it into a standard regression. Consider the case for a single currency.
If \([C_t/C_{t+1}, P_t/P_{t+1}, (F_t - S_{t+1})/S_t]\) are jointly log-normal then \(M_{t+1}\) is also log-normally distributed. Taking logs we can write the regression equation as

\[
\ln\left(\frac{F_t}{S_t}\right) + \ln\left(\frac{P_t}{P_{t+1}}\right) = -\rho \ln\left(\frac{C_t}{C_{t+1}}\right) + \ln M_{t+1} + \epsilon_{t+1} 
\]

(21)

The estimates under OLS would be inconsistent since both right-hand side regressors are correlated, but with instrumental variables, as in GMM, the estimates would regain consistency. This illustration demonstrates that the large estimate of \(\rho\) arises due to the high variability in excess returns, and the low variability in consumption growth. This then requires small changes in consumption to map into large changes in returns (a large coefficient) for a given R-squared and returns series. This puzzle is typically phrased as consumption not being variable enough to account for returns without implausible estimates of risk aversion. The models are estimated assuming the future values for the spot rate, prices, and consumption are known with perfect foresight (up to a random error). To emphasize again the ambiguity due to the joint hypothesis of REH though, it could be the issue is not consumption growth having too low of variability, but rather ex post returns are too variable of a proxy for ex ante returns.

### 3.5 A Generalization to Consumption Durability and Habit Persistence

Backus, Gregory and Telmer (1993) adapt a variant of the CCAPM to the context of the foreign exchange market.\(^{13}\) This model allows for the possibility of habit persistence (higher consumption in the previous period requires higher consumption this period to maintain the same utility) or durability (consumption in previous periods adds to utility in the current period). In their model, utility is not separable over time but is of the form of constant relative risk aversion

\[
U_t = E_t \left[ \sum \beta^k U(d_{t+k}) \right] 
\]

(22)

where \(d_t = c_t - \eta c_{t-1}\)

When \(\eta\) (referred to as the habit parameter, since it defines the intertemporal non-separability of utility) > 0 there is habit persistence, meaning higher

---

\(^{13}\)Early applications of this variant of the CCAPM to address the equity premium puzzle can be found in Nason (1988), Sundaresan (1989), Abel (1990), and Constantinides (1990). The excess returns puzzle with currencies though focuses not on the mean returns of one asset over another (e.g. stocks over bonds), but rather their fluctuations (e.g. changes in excess returns of one country’s bond over another).
current consumption requires higher future consumption to maintain future utility. When \( \eta < 0 \) consumption is durable such that current consumption adds to future utility. Their data set employs monthly U.S. dollar exchange rate data from July 1974 to April 1990 with the Canadian dollar, French franc, mark and yen. Consumption is measured as U.S. expenditure on nondurables and services, excluding clothing and medical care, and the implicit consumption deflator is used to measure prices. The forward discounts, and lags of consumption growth and inflation are used as instruments.

Backus, Gregory and Telmer follow Mark’s use of GMM to estimate the Euler condition which in this case is formulated as:

\[
E[\mu_{t+1}r_{t+1}] = 0
\]

(23)

where \( \mu_{t+1} = (1 - \beta \eta)^{-1}[\beta d_{t+1} - \beta^2 \eta d_{t+2}](p_t/p_{t+1})/d_t\)  

(24)

Backus, Gregory and Telmer first estimate the model imposing \( \eta = 0 \), corresponding to Mark’s earlier work, so \( d_t = c_t \). They again find very large estimates for \( \rho > 50 \). When they allow for persistent effects of consumption, they find that \( \eta < 0 \), implying a higher value for \( \rho \), which indeed is estimated at over 100.

In order to determine the source of the rejection, the authors calibrate a model attempting to select values for the parameters to correspond to the moments of the forward discount and returns data. The major difficulty it appears is in producing a model which is simultaneously compatible with the high serial correlation found in the forward discount, and the high variance of returns, the latter of which again could be a by-product of using ex post returns to proxy for ex ante.

As Engel (1996) concludes the overall deduction from the studies using consumption data is that the models do not work very well. Consumption generally does not vary enough to account for the high variation in ex post returns on foreign exchange without extremely large (and thus generally regarded as implausible) degrees of relative risk aversion. The models are also found to be inconsistent with the predictive power of the forward rate. Again though the tests of both the CAPM and CCAPM models are suspect, since the modeling generally assumes parameter stability, stationarity, and white-noise forecast errors, which could potentially invalidate any inference, and which have been ubiquitously rejected when explicitly tested\(^{14}\).

\(^{14}\)The case of the latter (tests of white-noise forecast errors) will be discussed in the following section.
4 Do we need an Alternative Model of the Risk Premium or Forecasting?

As with the previous tests of UIP, the difficulties encountered by the canonical models cannot be directly traced to the modeling of the ex ante component without further information, since the empirical work constitutes a joint hypothesis about both risk behavior \((\hat{r}_{t|t+1})\) and forecasting \((\varepsilon_{t+1})\). Recall the Bilson-Fama regression.

\[
\Delta s_{t+k} = \alpha + \beta f p^k_t + \varepsilon_{t+k}
\]  

(25)

In order to unravel the source of the forward discount bias, Froot and Frankel (1989) make use of survey data on traders’ actual exchange rate forecasts to proxy \(s^r_{t+1}\). This allows one to decompose the forward rate bias, or the deviation from the null hypothesis that \(\beta = 1\), into a component attributable to expectational errors, and another attributable to a risk premium. The probability limit of the coefficient \(\beta\) is

\[
\beta = \frac{\text{Cov}(\varepsilon_{t+k}, f p^k_t) + \text{Cov}(\Delta s^r_{t+k}, f p^k_t)}{\text{Var}(f p^k_t)}
\]

(26)

Where \(\Delta s^r_{t+k}\) is the expected depreciation, as measured by the median survey response, and \(\varepsilon_{t+k}\) is the expectational error of the median forecast. By taking the definition of the risk premium \(\hat{r}_{t|t+1} = f d^k_t - \Delta s^r_{t+k}\) and rewriting in terms of \(\beta\) equals one minus a term arising from the risk premium and one from a failure of REH. \(\beta = 1 - b_{rp} - b_{re}\) where

\[
b_{rp} = \frac{\text{Var}(\hat{r}_{t|t+1}) + \text{Cov}(\hat{r}_{t|t+1}, \Delta s^r_{t+k})}{\text{Var}(f p^k_t)}
\]

(27)

\[
b_{re} = -\frac{\text{Cov}(\varepsilon_{t+k}, f p^k_t)}{\text{Var}(f p^k_t)}
\]

(28)

With the use of survey data both of these terms are observable. A result that \(b_{re} = 0\) would suggest that forecast errors over the sample are uncorrelated with the forward discount, while \(b_{rp} = 0\) would suggest that the risk premium is uncorrelated with the forward discount. The magnitude of \(b_{re}\) is generally much larger, sometimes by an order of magnitude. Further, the sign of \(b_{rp}\) is in some cases negative, suggesting that the risk premium actually tends to push the bias in the opposite direction, to an estimate of beta larger than one. In order to evaluate whether these estimates are statistically significant, Froot and Frankel test a few competing hypotheses.
\[ \Delta s_{t+k} = \alpha + \beta f p^k t + \varepsilon_{t+k} \] (29)

The above equation represents a direct test of UIP, as opposed to the ex post tests which are more accurately a test of forward rate unbiasedness. The error term now is interpreted as the measurement error of the survey data (in proxying the relevant aggregate market belief). The null hypothesis that the risk premium is uncorrelated with the forward discount is \( \beta = 1 \), given \( \beta = 1 - b_{rp} \).

In order to fully test UIP \( \alpha = 0 \) would also have to hold. In seven of the nine cases, they cannot reject that the estimate of \( \beta = 1 \), implying that the risk premium is not statistically significant. In the other two cases, it appears that the risk premium accounts for more than 100% of the bias, and in the other the risk premium is negative indicating that it actually pushes the bias in the other direction, above one, accounting for a negative percentage of the bias.

Froot and Frankel consistently find a large constant term in the regression (rejecting the null of \( \alpha = 0 \) at less than .1% significance levels in all cases). This suggests that individuals are not risk neutral as implied by UIP. This term should not be interpreted as implying a constant risk premium however, but rather a large average level of the premium over the sample which in their study appears to be largely uncorrelated with the forward discount. If one wanted to test the proposition that the premium could be represented as a constant over the sample, tests for cointegration would be necessary to see if the deviations from equation (29) were stationary or persistent (this is conducted in chapter two).

In order to test whether prediction errors account for the forward bias, Froot and Frankel run the test of excessive speculation discussed by Bilson. This stringent test of REH asks whether investors would do better to place more or less weight on the spot rate rather than all other variables in their information set.\(^{15}\) The null hypothesis is that \( \alpha = 0, \beta = 0 \). The error term is the measurement error of the surveys minus the unexpected change in the spot rate. This allows one to test whether the estimates of \( b_{re} \) are statistically significant, since the coefficient \( \beta \) here is exactly equal to \( b_{re} \).

\[ \Delta s_{t+k} - \Delta s_{t+k} = \alpha + \beta f p^k t + + \varepsilon_{t+k} \] (30)

Indeed, in all cases \( b_{re} > 0 \) corresponding to excessive speculation, and is precisely estimated. Bachetta, Mertens and Van Wincoop (2006) also find that the predictability of returns in equities, foreign exchange, and bond markets appear related to expectational errors. Using survey data they find that the same variables which predict expectational errors, generally predict excess returns. They do not however examine the properties of the risk premium and its connection to excess returns.

\(^{15}\)Froot and Frankel (1985,1987) and Dominguez (1986) also test for biasedness in survey data.
4.1 Rethinking the Conclusion: Risk may Matter while Irrationality May Not

Some have used this seminal finding of Froot and Frankel to argue in favor of explanations for excess returns based solely on pure irrationality.\(^\text{16}\) By pure irrationality would be the view that there are obvious profits to be made from simply betting against the forward rate and market participants are merely unaware of this regularity, thereby perpetually forgoing obvious profit opportunities. If true this would represent a truly gross form of market inefficiency.

The first issue to note though is that Froot and Frankel pooled their data across exchange rates (for the same source and time horizon of the data), which implicitly assumes the relationship is the same for all exchange rates within each set. Other studies which use alternative samples, and unpooled data (using bilateral exchange rates) find evidence both of violations of REH (as do Froot and Frankel 1989) and a time-varying risk premium (Cavaglia, Verschoor, and Wolff 1994, and Chinn and Frankel 1994, 2000). When taken as a whole then, the survey data literature suggests that risk does matter.

As a demonstration of the obscuring effects of pooling the data across exchange rates, presented below are a subset of Cavaglia, Verschoor’s and Wolff’s results using bilateral data, which we can then contrast with the results using pooled data. The chi-squared statistic is the test of perfect substitutability, the joint hypothesis that \(\alpha = 0\) and \(\beta = 0\) in \(\Delta s_{t+k} = \alpha + \beta f_p^k + \varepsilon_{t+k}\). One can see that the bilateral results reject the null in three of four cases at the 5% level (indicated by the *) and in two cases at the 1% level (indicated by the **), and yet the pooled results do not.

<table>
<thead>
<tr>
<th>Exchange Rate</th>
<th>(\chi^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JY/DM</td>
<td>8.70*</td>
</tr>
<tr>
<td>SF/DM</td>
<td>0.66</td>
</tr>
<tr>
<td>BP/DM</td>
<td>91.49**</td>
</tr>
<tr>
<td>CD/DM</td>
<td>41.45**</td>
</tr>
<tr>
<td>Pooled</td>
<td>10.91</td>
</tr>
</tbody>
</table>

While these later results draw into question Froot and Frankel’s conclusion concerning the empirical relevance of a time-varying premium (or the lack thereof), they were very careful to note that their results only conclusively demonstrate that the common practice of using ex post data to infer about ex ante expectations is likely to be misleading. They acknowledge that their results may not imply "obvious irrationality" (agents systematically forgoing obvious profit, or even utility increasing opportunities). They remark that this result could be due to the well-known peso problem, the possibility of a large devaluation which simply has not occurred within the sample.

Another alternative they cite is that there could be changes in the process governing the spot rate, related to the parameter instability discussed previously.

\(^{16}\)See the section "Explanations based on non-REH forecasting and Microstructure" for further discussion of this literature.
with regard to the results on forward rate biasedness. Frydman and Goldberg (1996) and other studies find evidence of such structural change in the monetary models of the exchange rate. Frydman and Goldberg’s contingent market hypothesis argues that, given the openness of the world and non-routine change, market participants are likely to alter their forecasting strategies at least intermittently over time in ways that cannot be pre-specified in advance, which itself is likely to engender change in the process driving market outcomes. If the relationship between returns and the forward rate is changing over time, this would belie the interpretation that market participants are foregoing obvious profit opportunities to be made simply by betting against the forward rate. The survey data results implying irrationality have not explicitly tested for parameter stability, which is an important aspect of verifying the robustness of any empirical conclusion, and in turn must be regarded with a degree of skepticism. Indeed, a quite recent working paper, Frydman, Goldberg, and Kozlova (2013) finds evidence of parameter instability in the correlation between the forecast error and the forward rate, where it actually reverses sign in certain sub-samples. This evidence strongly refutes the interpretation of obvious irrationality.

Another issue is that the persistence of the variables has also not been suitably addressed, which may cause the results to be inconsistent. Moreover, even if UIP appears to "hold" ex ante (based on the survey data) with a coefficient indistinguishable from unity (which it actually does not once using unpooled data), the deviations from the estimated relationship may be persistent, meaning it should not be interpreted as an equilibrium relationship. Persistent deviations from the relationship would suggest that either the measurement error is not white-noise, and/or there is a time-varying (non-stationary) risk premium, though a coefficient of unity would imply that the time-varying risk premium could not account for the forward bias.

4.2 An Omission in the Literature

While the findings of non white noise forecast errors suggest a need to revise the specification of forecasting, and the presence of a risk premium suggests that risk does matter, the existing literature has not answered the question of whether any of the extant models can account for movements in the risk premium. In other words, they have not tested whether the traditional models can account for the risk preferences of actual traders which we have observed directly through the expected excess return in survey data. One possibility is that the time-varying risk premium, even if non-stationary, may simply be a persistent measurement error, in that we of course only have a sample of traders. Consequently the finding of a time-varying risk premium may not fully invalidate the hypothesis that relative returns equalize ex ante for the market in aggregate. If, however, we can connect the sampled risk premium to some risk factor, we can be much more confident in our interpretation that it represents a genuine risk premium.
5 Experimental Evidence on Risk Preferences

Given the finding of a time-varying risk premium, it is important to determine which model(s) can account for its behavior and to what extent. While the canonical models have not been tested on the basis of ex ante survey data to see whether they can account for the behavior of actual market participants, there is much experimental evidence that EUT is inconsistent with individuals’ preferences towards gambles. EUT, at least on its own, is thus liable to provide a faulty foundation for our models of the risk premium. A classic example of the type of risk averse behavior theory aims to explain was provided by Samuelson. An individual is given the option to accept or decline the following gamble: one which is equally likely to pay $110 or cost you $100, denoted as ($110, .5; -$100, .5), so a simple flip of the coin for example where heads pays you $110, and tails loses you $100. Despite the fact that this is a positive expected value gamble bet ($110 * .5 – $100 * .5 = $5), individuals will often forego the bet in favor of the lower expected value, but certain, payoff of zero.

It may seem surprising that we need to depart from EUT in order to explain such behavior. Rabin (2000) demonstrated however that we cannot interpret aversion to small positive EV bets in the framework of EUT without leading to implausible conclusions. He provided a non-parametric theorem that demonstrated that if an expected utility maximizer rejects the above gamble, they would also reject the gamble lose $1000 half of the time, and win any sum of money the other half of the time ($\infty, .5; -$1000, .5)! Certainly this seems like an implausible degree of risk aversion. The intuition of the result is quite simple, given a smooth, increasing, concave function in final wealth, if an individual is willing to forego the first bet, this must imply a high degree of risk aversion and an extremely concave utility function. In turn, given this high degree of concavity, the individual must be extremely risk averse to large potential losses, such as the loss of $1000.

There are other features of individual behavior towards gambles which contradict the predictions of EUT. A well-known example from Kahneman and Tversky’s work depends on two sets of decisions. In the first, assume in addition to whatever you own you are given $1000. Now you must choose between A) a guaranteed pay-off of $500, and B) a gamble which is equally likely to pay $1000, and $0 ($1000, .5; $0, .5). Individuals usually select option A) given this choice, which coincides with the standard understanding of risk averse behavior. In the second, set of options, now pretend that in addition to whatever you own that you are given $2000.

Now you must select between C) a guaranteed loss of $500 and D) a gamble which is equally likely to lose $1000 and $0 ($-1000, .5; $0, .5). In this case however individuals often select option D, which would imply risk-seeking behavior under the standard interpretation of EUT. Note that the two sets of options (A and B, and C and D) are identical in terms of the levels of final (expected) wealth, and yet individuals often select differently. This example, when combined with Samuelson’s example, highlights the type of behavior Prospect
Theory attempts to formalize in as parsimoniously as possible.

It appears that the framing of the question, whether in terms of gains or losses, alters the response. This suggests that individuals are reference dependent, evaluating outcomes in terms of gains or losses relative to some reference level, typically discussed as the status quo or initial level of wealth though the exact reference level may depend on the context. This is in contrast to expected utility theory, where the final level of wealth is all that matters in the evaluation of utility. Of course, a simple example would suggest that two people, even with identical preferences and wealth, may not have the same utility. Consider two individuals with $1,000,000 in wealth, if for example one individual previously had wealth of $100,000 and the other wealth of $10,000,000, they would likely have quite differing levels of contentment from now possessing $1,000,000, the former having experienced a large gain likely being quite ecstatic, while the former having experienced a large loss may be quite disgruntled.

This type of perception in terms of changes from recent conditions, rather than in terms of absolute levels, is actually a very natural by-product of adaptation and is consistent with the way we perceive a wide range of phenomena. Consider for instance the way you respond to temperature, sound, or brightness. The second feature of Prospect Theory is designed to capture aversion to small expected value bets, coined loss aversion, whereby the disutility of a loss is greater than the utility from an equal magnitude gain, usually on the order of greater than 2:1. The last feature of prospect theory is diminishing sensitivity to both gains and losses, which would explain the finding that individuals were risk averse over large gains (since a gain of $1000 is not twice as valued as a gain of $500) and risk seeking over large losses (since the loss of $1000 would be less than twice as painful as the loss of $500). These features can be represented with a Prospect Utility function, called the value function as in Kahneman and Tversky (1979, 1992). We have the value depending on gains and losses, in a piece wise fashion with a larger weight on losses, and concavity in gains and convexity in losses.

\[
v(\Delta W) = (\Delta W)^\alpha \quad \text{if} \quad \Delta W \geq 0
\]
\[
v(\Delta W) = -\lambda(\Delta W)^\beta \quad \text{if} \quad \Delta W < 0 \quad \text{where} \quad \lambda > 1 \quad \text{implies loss aversion}
\]

In formulating the utility function representation of their experimental results, Kahneman and Tversky assume that the exponents on the change in wealth (the curvatures) are equal in both gains and losses ($\alpha = \beta$), though the results could also be consistent with ($\alpha < \beta$).

Recall that in the work of Rabin (2000) interpreting behavior towards gambles in light of EUT led to implausible degrees of risk aversion. Similarly, attempts to explain the premium in asset markets in light of EUT leads to implausible degrees of risk aversion. Prospect theory meanwhile is able to provide an explanation for the experimental evidence on individuals’ behavior towards risk. Perhaps then applying prospect theory will be able to help in resolving the excess returns puzzle. A major difficulty in applying prospect theory to financial
markets is in specifying how individuals forecast the potential loss, as opposed to the experimental settings where this is given, in addition to the previously emphasized difficulty of representing how individuals form their point forecasts of the expected return.

6 Alternative Models of the Premium

Several models of the premium have been developed on the basis of prospective utility preferences, in conjunction with other specifications of decision-making in financial markets. Barberis, Huang and Santos (2000), hereafter BHS, use a mixed approach of both EUT and loss aversion to match the historical premium on equities over government bonds, and other stylized facts including the volatility and predictability of returns.\footnote{BHS draws from the earlier application of loss aversion by Benartzi and Thaler (1995) to create a dynamic model.} This suggests a potential benefit from integrating EUT and prospect theory. They also use another finding from psychology concerning the way individuals’ preferences towards risk evolve over time based on previous performance. It has been found that prior gains tend to cushion the pain of subsequent losses, producing less loss averse behavior and vice versa for previous losses (Thaler and Johnson 1990).

BHS depart from the CCAPM by assuming that individuals get utility not just from consumption, but also from the gain or loss in wealth. In turn, the representative agent will maximize:

$$E[\sum (\beta^{C1-\rho} + b_t \rho^{t+1} v(\Delta W_{t+1}))]$$

(32)

Where all symbols are as before, though $\Delta W_{t+1}$ is related to the return on risky investments, and $b_t$ is a scaling factor proportional to aggregate per capita consumption, which determines the weight attached to investment aside from its impact on consumption. Returns are assumed to be evaluated relative to the reference level of the risk free rate. For tractability, BHS ignore the non-linearity of prospect theory, which they justify by noting that the results from Benartzi and Thaler (1995) find that most of Prospect Theories ability to match the excess returns puzzle in equities comes from loss aversion, while the non-linearities appear to be rather unimportant to the results.

$$v(\Delta W) = \Delta W \text{ if } \Delta W \geq 0$$

$$v(\Delta W) = -\lambda \Delta W \text{ if } \Delta W < 0 \text{ where } \lambda > 1 \text{ implies loss aversion}$$

(33)
The model also employs both loss aversion and risk aversion, in order to account for limits to speculation, which suggests that perhaps both effects are of relevance for understanding returns.

The initial empirical work on the BHS house-money effect model was based on REH and calibration, using the prospect theory estimates of $\lambda = 2.25$, in an attempt to match the historical excess return of equities over government bonds. The literature on the analogous puzzle in the foreign exchange market however is focused instead on changes in excess returns over time, and as such it would be desirable to evaluate this model in the foreign exchange market on the basis of its ability to match the time path of excess returns.

A key feature of the model is that the degree of loss aversion depends on previous performance, such that a previous gain reduces loss aversion while vice versa for a gain. Without this BHS are unable to generate the sizable premium typically observed, regardless of the evaluation period used.

The importance of the evaluation period is that, given equity prices are more variable, a shorter time horizon would imply more observations of losses over a given period, which are weighted more heavily given loss aversion, and thus a lower prospective utility over any given period the more frequent the evaluation of gains and losses. Benartzi and Thaler (1995) estimated an evaluation period of approximately one year, which they argue is consistent with tax filings and the fact various funds and pensions provide their most comprehensive reports on an annual basis. In order to generate a large premium, returns need an additional source of volatility. The dynamic loss aversion implies that when dividends increase, and the price increases, loss aversion decreases causing an additional increase in the price. In order to model the effect of prior outcomes, they introduce a benchmark price, here referred to as $S_{bm}^{t}$, which can be thought of as a second reference level. The only restriction placed on $S_{bm}^{t}$ is that it moves more gradually than the price itself (less than one-for-one). Loss aversion then is related to deviation of the price $S_{t}$ from its benchmark level, specified in ratio form $S_{t}/S_{bm}^{t}$. By calibrating the parameter relating loss aversion inversely to prior outcomes they are able to match the volatility of equity prices and the mean historical premium, the high volatility of equity prices, and the predictability of returns based on price-to-dividends ratios.

A second model which uses a variant of PT comes from Frydman and Goldberg (2003, 2007), hereafter FG, called the Keynes-imperfect knowledge economics (IKE) gap model. Recall that in Kahlman and Tversky’s formulation, it was assumed that the curvature in gains and losses was equal. In general though, the degree of loss aversion would be

$$\Lambda = \lambda(\text{W} | \text{ar}^l |)^{\beta-\alpha}$$  \hspace{1cm} (34)

Where $\text{W} | \text{ar}^l$ is the position size and $\text{ar}^l$ is the forecasted potential loss (the mean from the loss portion of the forecasted distribution of returns) on a unit position. Only in the special case where $\beta = \alpha$ does the degree of loss aversion
reduce to the constant $\lambda$. The experimental results however could also be consistent with $\beta > \alpha$. In this case, for a given $r^l$, the degree of loss aversion will increase with the size of one’s position. Since individuals select their position size, FG refer to this as endogenous loss aversion. In order to ensure that $\Lambda > 1$ over the domain of losses, they alter the representation of the degree of loss aversion to be:

\[ \Lambda = \lambda_1 + \lambda_2 (W|r^l|)^{\beta - \alpha} \]  

where $\lambda_1 > 1$ and $\lambda_2 > 0$ are constant preference parameters. Given $\lambda_1 > 1$, $\Lambda$ is ensured to be greater than one. With this equation for endogenous loss aversion, the utility function becomes:

\[ V(\Delta W) = \begin{cases} (W|r^g|)^\alpha & \text{if } \Delta W > 0 \\ -\lambda (W|r^l|)^\beta & \text{if } \Delta W < 0 \end{cases} \]  

where $r^g$ denotes the potential unit gain. This representation is consistent with the three experimental results discussed previously, reference dependence, loss aversion, and diminishing sensitivity.

This resultant decision rule implies some important aspects of behavior. The first is that even when an individual perceives a gain from speculation, they will take a position of limited size due to endogenous prospect theory. Another is that individuals may perceive a profit opportunity, but opt to hold no position if the expected return is insufficient.

From the momentary equilibrium, and portfolio balance the basic equilibrium condition is again

\[ \hat{r}_{t|t+1}^l = \rho \hat{r}_{t|t+1} \]  

Here however $\hat{r}_{t|t+1}$ is a weighted average of the forecast of bulls holding a long position, and bears holding short positions. FG follow Delong et. al. (1990) in assuming that there is no reason to expect the wealth of bulls to exceed the wealth of bears or vice versa so

\[ \hat{r}_{t|t+1}^l = \frac{1}{2}(\hat{r}_{t|t+1}^L - \hat{r}_{t|t+1}^S), \text{ and } \hat{p}_{t|t+1}^l = u \hat{p}_{t|t+1} + \lambda_2 IP \]  

The uncertainty premium again is an equal weighted average of bulls and bears

\[ u \hat{p}_{t|t+1} = \frac{1}{2}(u \hat{p}_{t|t+1}^L - u \hat{p}_{t|t+1}^S) = \frac{1}{2}(1 - \lambda_1)(\hat{r}_{t|t+1}^L - \hat{r}_{t|t+1}^S) \]  

\(^{18}\)For further details on the derivation and constraints imposed see Frydman and Goldberg (2007) Chapter 9.
The remaining question then is how market participants forecast the potential unit loss from speculation. FG model this uncertainty premium as a function of the deviation of the asset’s price from its (more stable) benchmark level. This notion that risk depends on the deviation of an asset’s price from estimates of its historical or fundamental benchmark level draws on a lost insight of Keynes (1936), which FG refer to as the gap effect. As the price (or forecasts of the future price) move further away from estimates of its benchmark value, bulls will become more worried about the potential for and possible magnitude of a reversal, and will increase their forecast of the potential for loss. Consequently, they will demand a greater uncertainty premium to increase or maintain their long positions. Meanwhile bears will become less concerned about a further movement away, lowering their forecast potential loss and requiring a lower premium. Both the effects for bulls and bears will act to increase the aggregate uncertainty premium. As in the CAPM, the premium depends on the net long position the market must hold (\(IFP_t\)), but rather than being related to the variance of the exchange rate, it is related to this gap effect. DeGrauwe and Grimaldi (2006) similarly model the degree of risk aversion in certain versions of their models as depending on the difference between the exchange rate and its fundamental value. In their model chartists extrapolate past price movements, and fundamentalists expect a return to some fundamental level. In the case of the latter, as the exchange rate moves further from the fundamental value, they become less risk averse. Beyond explaining the time path, this gives the model the opportunity for more frequent sign reversals, aside from a switch in the IFP term or nations alternating from net debtor to net creditor, which could potentially remedy the sign reversal puzzle discussed in section 3.2.

In the case of foreign exchange, a natural candidate for such a benchmark value is the purchasing power parity (PPP) exchange rate, which is the level of the exchange rate which equalizes costs across countries. There is a large literature on mean reversion of the exchange rate back to this level (Rogoff 1996, Taylor and Taylor 1996). To be quite concrete about what is meant by the gap effect, the below three graphics show the spot exchange rate in black, and an estimate of PPP in blue (based on the Economist’s PPP Big Mac Index and extrapolated with consumer price indices). The gap then is the difference between the two series. The graphics are quite consistent with the stylized facts of floating exchange rates, where the exchange rate tends to fluctuate in long swings of irregular size and duration around the much more gradually moving benchmark of PPP. Ultimately though there is a tendency to mean revert back to that level, at least in some non-linear fashion, yet there is no tendency to stabilize around it, since conditional on its return it often continues trending in the same direction.
The equilibrium condition implies a cointegrating relationship between the premium and the real exchange (gap). FG refer to this equilibrium condition as Uncertainty-adjusted UIP (UAUIP)

This relationship is consistent with other findings that the premium predicts a negative skewness to returns, or currency crashes (large adverse movements in the exchange rate.) Stated differently, the currency which garners a premium possesses more "downside risk" i.e. is more likely to undergo a steep and sudden decline in value. Although crashes are not frequent or large enough to equalize returns, this can be rationalized if the negative "peso states" are weighted more heavily (Brunnermeier, Negal and Pedersen 2008). This greater weighting of the peso states (losses) is consistent with loss averse behavior. The model of Frydman and Goldberg then provides additional insight to this observation by suggesting how individuals forecast the potential for currency crashes, which leads them to demand a premium in the first place. Specifically it is hypothesized that investors do so by looking to the gap between the exchange rate and its historical or fundamental benchmark level.

6.1 Conflicting Predictions of the House Money and Gap Effects

Despite the fact both the house money and gap models rely on prospective utility preferences, they actually generate contradictory predictions. The house money effect would predict that the premium should co-vary negatively with upswings of the exchange rate (Frydman and Goldberg 2007 note this implication of the house money model), as the exchange increases loss aversion decreases and thus so does the equilibrium premium. The IKE gap model meanwhile predicts that the premium should co-vary positively with upswings of the exchange rate, as the exchange rate increases the forecast of potential loss increases, and thus the equilibrium premium. As is clear, the two models imply opposing qualitative
predictions about the co-movement between the premium and the asset price. The two models can be directly related if one assumes the same benchmark in both models.

Frydman and Goldberg (2007) use survey data to measure the premium, and test the prediction of these two models.

\[ \hat{r}_{t|t+1} = \beta(s_t - s_{t}^{PPP}) \]  

(40)

If one assumes the benchmark in both cases is the PPP exchange rate, testing can be done with the real exchange rate as in Frydman and Goldberg (2007).\(^{19}\) A coefficient of \(\beta < 0\) would support the house-money effect dominating (as the exchange rate increases, bulls reap a gain, and the house-money effect leads to a decrease in loss aversion and the premium). If \(\beta > 0\) this would imply the gap effect dominates (as the exchange rate increases, the forecast of potential loss increases, and the premium increases). Frydman and Goldberg find statistically significant estimates of \(\beta > 0\) for all three exchange rate samples they investigate.

7 Ways to Improve Testing

While the house money and gap effects are mutually exclusive, in that only one at most can be observed, it is possible that the effects of risk aversion (volatility or the covariance with consumption) may operate simultaneously with the dominant of these two effects. Previous studies however have not tested the ability of the canonical models to explain the premium as measured by survey data, nor do they allow for the effects of volatility or consumption to act in conjunction with the gap effect. This dual incorporation is motivated in part by BHS’s mixed model combining EUT and loss aversion.

In addition to this desirable expansion, the early work on the gap effect could be expanded and improved upon in a few ways on statistical grounds. Frydman and Goldberg estimate the model using a standard error-correction model to account for non-stationarity in the data (see Granger 1983 and Engel and Granger 1988). They do not however explicitly test for the non-stationarity of the variables, nor do they explicitly test for cointegration between the variables. More recent developments in the cointegration literature have advocated conducting estimation as a system of simultaneous cointegrating relationships, known as the Johansen method (Johansen 1988, 1995) or the Cointegrated VAR model (Juselius 2006). This allows the data to "speak freely" about the rank of the information set (number of relationships), and the stochastic trends and

\(^{19}\)Frydman and Goldberg (2007) use the Economist’s Big Mac PPP index, which tracks the cost of a McDonald’s big mac across countries to provide an estimate of price levels.
One major advantage of Frydman and Goldberg’s empirical work (Goldberg and Frydman 1996, Frydman and Goldberg 2007) is that they test for parameter instability in the relationship, without pre-specifying the placement or number of breaks, which is surprisingly uncommon in the literature. Typically empirical work just assumes parameter constancy. It would be useful however to extend the tests for structural change to newer alternatives designed specifically for non-stationary data (for example the eigenvalue fluctuation test of Hansen and Johansen 1999).

More recent work also suggests that the exchange rate and real exchange rate are more persistent than is commonly recognized. This work highlights the importance of conducting testing for persistent changes in the variables (near I(2) trends) using multivariate tests. This I(2) behavior is a generalization of the earlier broken trends model of Engel and Hamilton (1990) where the time-varying drift is stochastic rather than following a markov process between two states with opposite sign. A failure to fully address these large roots is likely to lead to spurious inference.

While the focus here is on explaining the ex ante premium, it would be illuminating to decompose the premium into its components (the interest rates, spot exchange rate, and expected future spot exchange rate). This would circumvent the potential endogeneity bias arising from having the same variable on both sides of the regression, which potentially invalidates the inference of a positive relationship between the premium and gap. The use of a composite premium variable also represents an untested assumption which ideally should be explicitly tested. Further, decomposing the premium would provide a clearer picture of the driving and adjustment process. Using a composite variable of the premium, as done in Frydman and Goldberg (2007), is liable to muddle the dynamics as it is likely some variables within it are exogenous while others are adjusting.

The I(2) model itself also allows for a much richer understanding of the driving and adjustment process. It identifies which variables produce the I(2) trends, their shocks leading to the persistent changes of the exchange rate. It also allows for a multi-tier adjustment to long-run relationships (the changes adjusting to the relationship in levels as in the standard error-correction model) and a medium-run adjustment (with the acceleration rates adjusting to the changes in other variables). This allows the empirical model to capture more naturally the "boom-bust" type processes or long swings often observed in asset markets, manifesting as error-increasing behavior where the price moves away from equilibrium in the medium-run, which the traditional error-correction model is inherently less able to capture.

In summary of the insights for future research: 1) One source of the difficulty experienced by the CAPM may be that it simply has the wrong measure error-correcting forces in the system (which variables are weakly exogenous and endogenous respectively).20

20The phrase "speak freely" is borrowed from Hoover, Johasen and Juselius (2007) which outlines the CVAR methodology.
of volatility. Using a realized measure may rectify this, as they have encountered superior performance relative to GARCH formulations in other contexts.

2) There is reason to believe however that one should begin by testing premium models against the ex ante survey measures, rather than the ex post measures which have been widely determined to yield a misleading proxy for the former.

3) Tests conducted with survey data should explicitly test for non-stationarity to determine whether the risk premium is truly a constant, and whether the estimates are consistent.

4) The instability of the BF regression suggests that explanations for the forward discount based on misforecasting and misforecasting about fundamentals should likewise test for parameter constancy. This has major implications for the deduction that the bias is due to obvious irrationality.

5) The experimental evidence suggests that EUT is a dubious assumption upon which to base a risk premium model. Models based on PT are a promising avenue to address the excess returns puzzle. The work of Barberis, Huang and Santos (2001) suggests that both risk aversion and loss aversion may be important, suggesting a role to simultaneously estimate effects of the canonical and newer alternative models of the premium.

6) Such testing should be conducted as a system using more powerful multivariate tests for stationarity.

7) Use of the I(2) model, and decomposing the premium into its components, would eliminate endogeneity bias and allow for more rigorous testing and a richer understanding of the driving and adjustment process in the foreign exchange market; including the source of the additional persistence in the real exchange rate (the I(2) trends) and the nature of the boom-bust process often observed (potential error-increasing behavior).

8 Explanations based on non-REH forecasting and Market Microstructure

The difficulty of the canonical risk premium models to account for returns, and the early results using survey data which suggest that the forward rate bias is due primarily to violations of REH have led some researchers to develop models which depart from REH. Some of these models assume risk neutrality as a matter of convenience, and a desire to focus on what is novel in their work (Gourinchas and Tornell 2004). It is worth emphasizing however that if one is willing to attribute the forward discount not only wholly, but even in part, to a lasting, systematic error in forecasting on the part of the aggregate market, that one is saying an investor could beat the market and make excess profits merely by following a rule as simple as betting against the forward rate. This would constitute a truly gross form of market inefficiency.

None the less, it does seem that the development of alternative representations of forecasting is a worthy line of research, particularly based on the results of the survey data, if we are attempting to represent real world behavior. One should be a bit skeptical of the notion that market forecasting errors follow a fixed stochastic process, let alone one which is purely white-noise.
The existing behavioral models generally pursue two avenues to depart from REH. The first maintains a representative agent, and attributes the forward bias to systematic forecasting errors about the process driving fundamentals. Gourinchas and Tornell (2004) for example relate the forward discount to systematic underestimates of the persistence of interest rate shocks. They find evidence, based on survey data of traders’ forecasts about interest rates, that there does seem to be a relation between the forecast error about interest rates and the forward discount. Burnside et. al (2011) similarly relate the bias to misperceptions of the variance of a fundamental, in this case prices. Their model represents agents as underestimating the variance of their predictions and thus excessively speculating, which they refer to as over-confidence. In both cases, the authors interpret their results to imply that there are obvious profit opportunities to be gained, simply by betting against the forward rate (though these are said to be rather small). Bachetta, Mertens and van Wincoop (2007) also find based on survey data that the predictability of returns in several markets is related to forecasting errors, though in later work Bachetta and van Wincoop (2011) explain this through models of rational inattention. Gourinchas and Tornell (2004) similarly note that their findings could be the product of robust control behavior, where individuals worry about misspecification of their forecasting model, rather than naive mis-forecasting.

The second approach allows for heterogeneous expectations, typically of two forms, fundamentalists and chartists (or technical traders). The earliest model of this type was developed by Frankel and Froot (1989) to explain the USD/DM "bubble" in the mid-1980’s (a large appreciation followed by an essentially equal depreciation). DeLong et al. (1990) develop another model of heterogeneous expectations in the setting of the stock market. The model is populated by rational traders (fundamentalists) and "noise traders" who respond to erroneous information. The risk that noise traders may push the price further away from fundamental values creates a risk for the rational traders (given margin requirements and credit constraints) which impedes perfect arbitrage back to fundamental values. Mark and Wu (1998) develop this model in the foreign exchange market and argue there is preliminary empirical support for it based on survey data.

De Grauwe and Grimaldi develop a model with technical traders and fundamentalists, where individuals switch between the two strategies according to their relative past profitability, and a behavioral rule dictating the speed of switching (the length of time over which individuals compare profitability). The models are analyzed in an agent-based simulation, and are able to replicate many of the stylized facts, including volatility clustering, fat tails, long swings, a disconnect from fundamentals, and the profitability (and thus ubiquitousness) of technical trading. Thus De Grauwe and Grimaldi are not viewing the technical traders as clearly irrational (foregoing obvious profit opportunities) but rather view such adaptive forecasting strategies as the best one can do in a very complex world.

Another explanation for the excess returns puzzle focuses on the market microstructure, and some distortions which may arise due to commonly observed
trading practices in real world markets. Burnside, Eichenbaum, Kleshchelski, and Rebelo (2006) examine the returns to currency speculation which attempts to exploit the excess returns puzzle. They find that the trade is on average quite profitable (a large Sharpe ratio). They focus however on the practical implementation of the trade and transaction costs. They observe that the bid-ask spread is often observed to be an increasing function of order size. Another important facet of this literature is "price pressure", whereby the exchange rate is an increasing function of net order flow. This phenomena can be motivated by either asymmetric information and a response against adverse selection (Kyle 1985, and Easley and O’Hara 1987), or inventory motives (Garman 1976, Stoll 1978, and Cao, Evans and Lyons 2006). Evans and Lyons (2002) find strong empirical support for the price pressure phenomena. Burnside et. al. argue that these frictions greatly reduce the profitability of attempting to exploit the forward discount anomaly, and demonstrate that the marginal profitability (Sharpe ratio) can be zero even if the average is large. While this explanation can account for the lasting effect of this bias, it does not provide an explanation for why it arises in the first place. Burnside, Eichenbaum and Rebelo (2006b) however develop a model where the adverse selection problem is limited when a currency is expected to appreciate, which could account for the development of the bias.

The more recent findings however which suggest the presence of a time-varying risk premium as measured by survey, and one which appears to be correlated with the "gap" risk factor, suggests that these explanations, while seemingly part of the story, are at least incomplete, and indeed risk does matter. Further the findings of structural change undermine the view that individuals are foregoing obvious profit opportunities and the view that any given type of systematic misforecasting is occurring endlessly.

9 REFERENCES


Research Unit, Institute of Economics, University of Copenhagen.


