Predicting Commodity-Futures Basis Factor Return by Basis Spread

Daehwan Kim

Department of Economics, Konkuk University, Seoul, South Korea

dkim@konkuk.ac.kr

Abstract

A growing body of literature confirms the significance of the commodity futures basis factor: It has a significantly positive premium and it explains the cross-section of commodity-futures excess returns. We extend the literature by documenting predictive relation between this factor and the inter-quartile spread in the basis. Using commodity futures market data between 1972 and 2011, we show that the basis spread is a strong predictor of the basis factor return. Our finding supports the insight from recent theoretical models that economy-wide production shock affects the commodity market risk premium through the basis.

Keywords:

Commodity futures, basis factor, basis spread, predictability, risk premium

JEL Classification:

G10, G12, G13

Acknowledgement

The paper has benefited from useful comments of Sangwon Suh, Kyuhyong Kim, Wenjie Ma, Ke Tang, Jinho Bae, and other participants of the 2012 Korea International Economic Association Conference, 2013 Joint Conference of Finance Associations of Korea, 2013 Asia Pacific Association of Derivatives Conference in Busan, and the 2nd International Conference on Futures and Derivative Market in Beijing. This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012S1A5A8024387).

1. Introduction

A growing body of literature confirms the significance of the commodity-futures basis factor: It has a significantly positive premium and it explains the cross-section of commodity-futures excess returns. The basis factor is calculated as the return to the basis-sorted long-short portfolio, i.e. it is the average return of high-basis-commodity futures minus the average return of low-basis-commodity futures. Gorton and Rouwenhorst (2006), Fuertes, Miffre, and Rallis (2010), and Gorton, Hayashi, and Rouwenhorst (2012) report a significantly positive premium accruing to the basis factor. De Roon and Szymanowska (2010), Yang (2013), Szymanowska, de Roon, Nijman, and Goorbergh (2013), and Bakshi, Gao, and Rossi (2014) report that the basis factor explains a significant part of the cross-section of commodity-futures excess returns.[1]

We extend the literature by documenting the predictability of the basis factor. We calculate the inter-quartile spread in the basis from the cross-section of commodities, and show that this spread predicts the basis factor return.[2] When the next-period basis factor return (*y*) is regressed on the current-period basis spread (*x*), the coefficient estimate is statistically significant. An identity involving the basis and the excess return allows us to interpret the regression coefficient as the share of the *y*-related variation

[1] Szymanowska et al. (2013) use the term “the basis factor.” Yang (2013) calls it “the slope factor,” and Bakshi et al. (2014) refer to it as “the carry factor.” As noted by Szymanowska et al. (2013), the commodity-futures basis factor is comparable to the foreign-exchange-market carry factor; thus, the carry factor is certainly another suitable name for the basis factor.

[2] We calculate the inter-quartile spread in the *inverse* basis rather than in the basis, as the former is a more accurate representation of the theoretical relationship that we discuss later. Thus, the proper name for our spread variable should be the *inverse* basis spread. As it sounds rather clumsy, however, we call our spread variable simply the basis spread. Note also that our usage of this term is different from Yang’s (2013). In Yang’s paper, the basis spread refers to the premium accruing to the basis factor.

in the variation of *x*.[3] According to this interpretation, over 50% of the time variation in the basis spread is related to the variation in the next-period basis factor return.

The use of the basis spread as a predictor can be motivated in two ways. The first motivation comes from the positive cross-sectional correlation between the basis and the next-period excess return. All the studies cited above find that such relationship exists. Given this relationship, we may apply the "inter-quartile-spread operator" to each variable and obtain the positive relationship between the basis *spread* and the spread in the next-period excess return. The excess return spread is essentially identical to the basis factor return. Thus, the positive relationship will also exist between the basis spread and the basis factor return.

Our predictive regression is an application of a well known idea in the finance literature:
The return to a characteristics-sorted long-short portfolio can be predicted by the spread in the characteristics. This idea has been successfully applied to the prediction of stock-market factors, such as the value factor and the momentum factor. See, for example, Asness, Friedman, Krail, and Liew (2000), Cohen, Polk, and Vuolteenaho (2003), and Stivers and Sun (2010). As far as we known, this idea has not been applied to the commodity futures market factor, and that is what we attempt to do in this paper. That is, one of our contributions to the literature is to show that the spread-predict-the-factor-return idea can be applied to the commodity futures market.

The second motivation for our predictive regression comes from recent theoretical models of futures prices by Yang (2013) and by Gorton et al. (2012).[4] A common theme in these models is that both the

[3] Note that we are not referring to the R squared which is the share of the *x*-related variation in the variation of *y*.

[4] The model by Gorton et al. (2012) describes the economy with single commodity. We extend the model to the case of multiple commodities in Appendix A2 so that we can derive the implication on the basis spread. However, the main components of the implication exist in the original model.

risk premium and the basis are driven by the economy-wide production shock. Cross-sectional variation is attributable to commodity characteristics. Thus, if we take the spread in the basis, we remove the effect of commodity characteristics from the basis, and the resulting variable reveals the magnitude of the production shock. The same is true for the risk premium: The spread in the risk premium is proportional to the production shock. Therefore, both the basis spread and the risk premium spread move with the production shock, and the basis spread is a good predictor of the common factor return, which is a noisy proxy of the risk premium spread. This implication of the models has not been examined in the literature yet; so that is another area where our paper makes contribution.

The spread in other characteristics may also be useful as predictors. We consider the spreads in four other characteristics--the past return, the implied volatility, the hedging pressure, and the market interest. These characteristics have been examined by many authors in relation to commodity futures excess returns: See Miffre and Rallis (2007), Shen, Szarkmary, and Sharma (2007), and Fuertes et al. (2010) for the discussion of past returns; Gorton et al. (2012) for the implied volatility and other variables; Carter, Rausser, and Schmitz (1983) and Bessembinder (1992) for the hedging pressure; and Hong and Yogo (2012) for the market interest. Our analysis indicates that, among these four variables, only the spread in the past return is a highly significant predictor of the basis factor return. The predictive power of the past return spread is not surprising. The theoretical models that we have described above suggest that the past return spread is proportional to the past value of the production shock. Thus, to the extent that the production shock is serially correlated, the past return spread is related to the current production shock, and has the predictive power. This idea is consistent with the argument of Stivers and Sun (2010) that the cross-sectional dispersion in asset returns reveals the current state of the economy. We also consider the business cycle and predictors based on interest rates and bond yields. While the business cycle does not help the prediction, some interest-rate-based predictors have predictive power. In any case, the predictive power of the basis spread is not diminished by the inclusion of other predictors.

The remainder of this paper is organized as follows. In Section 2, we present theoretical motivations for our predictive regression. Section 3 discusses the predictive regression result. Section 4 examines alternative predictors, and we conclude in Section 5. The appendix has three parts: An algebraic derivation of the predictive regression equation, Eq. (3), is included in Appendix A1; an extension of the model by Gorton et al. (2012) is presented in Appendix A2; finally, data details are described in Appendix A3.

2. Motivation for the Predictive Regression

Why do we expect the basis spread to predict the basis factor return? One motivation comes from the empirical finding that the excess return is positively related to the basis; since the factor return is calculated as the difference between two portfolio returns, it might be related to the difference between portfolio bases. The basis spread is one measure of such difference. Another motivation comes from recent theoretical models which suggest that both the excess returns and the basis are driven by the economy-wide production shock; in these models, the basis spread reveal the magnitude of the shock and thus is a good predictor for the factor return. Below we review these two motivations more carefully.

Implication of Positive Correlation between Excess Return and Basis

The positive relationship between the excess return and the basis can be thought of in terms of the following "near" identity:

$$\frac{F_{i,t}}{S_{i,t}} - 1 \approx \left(\frac{S_{i,t+1}}{S_{i,t}} - 1\right) - \left(\frac{S_{i,t+1}}{F_{i,t}} - 1\right) \tag{1}$$

$S_{i,t}$ and $F_{i,t}$ are the spot and futures prices of commodity i at time t, respectively. If the futures contract expires at time $t+1$, $S_{i,t+1}$ can be viewed as the spot price at maturity. Then, $F_{i,t}/S_{i,t}-1$ is the "inverse basis"; $S_{i,t+1}/S_{i,t}-1$ is the "spot return"; and $S_{i,t+1}/F_{i,t}-1$ is the futures excess return.[5] That the "near" identity holds can be seen in the following way: if we replace each variable x with its logarithm $\log(1+x)$, then we obtain the exact identity. As each variable is fairly small, $\log(1+x)$ is approximately x, and we have the near identity.[6] We have chosen to express Eq. (1) in terms of non-logarithmic variables since later analysis involves cross-sectional averaging.[7] Let us denote the inverse basis by IB, the spot return by SR, and the excess return by XR. Then Eq. (1) can be written as:

$$IB_{i,t} \approx SR_{i,t} - XR_{i,t+1} \tag{2}$$

Eq. (2) merely states the near-identity relationship among variables, and does not reveal much on its own. In particular, Eq. (2) alone does not suggest the correlation between the inverse basis and the excess return. Now, let us suppose that the correlation between the inverse basis and the spot return is less than perfect. If that is the case, then we are guaranteed to have some correlation between the inverse basis and the excess return. Indeed, in the real data, the correlation between the inverse basis and the spot return is less than perfect. In the FX market, Hansen and Hodrick (1980) and many other authors have found that the forward price is not a good predictor of the next-period spot price, and thus the inverse basis is not perfectly correlated with the spot return. Fama and French (1987) and

[5] We define the excess return of commodity futures as the return to the investor who takes a long position in one futures contract, assuming that the investor's equity equals the futures price. Note that the investor's equity does not have to equal the futures price. The required margin for a futures position is typically much lower than this. Making alternative assumptions on the equity size leads to alternative definition of excess returns.

[6] As can be seen in Table A1, the mean values of the excess return and the inverse basis are less than 0.01. So the approximation error is less than 0.005.

[7] The cross-sectional average of non-logarithmic returns can be interpreted as a portfolio return, whereas the cross-sectional average of log returns does not have such interpretation.

others have found the same pattern in the commodity futures market. Given the less-than-perfect correlation between the inverse basis and the spot return, the correlation between the inverse basis and the next-period excess return is to be expected. A growing literature examines this relationship. See, for example, Gorton and Rouwenhorst (2006), Fuertes et al. (2010), and Gorton et al. (2012).

Our predictive regression can be motivated from Eq. (2) and the positive relationship between $IB_{i,t}$ and $SR_{i,t}$. Suppose that the relationship between $IB_{i,t}$ and $SR_{i,t}$ and also the relationship between $IB_{i,t}$ and $XR_{i,t+1}$ are monotonic. In Appendix A1, we show that the basic structure of Eq. (2) remains when we replace each variable with its inter-quartile spread, so that

$$\Delta IB_t \approx \Delta SR_t + \Delta XR_{t+1} \tag{3}$$

In the above, Δ indicates the inter-quartile spread--the difference between the 1st and the 3rd quartiles--or the spread based on some other quantiles. Note the positive sign in front of ΔXR_{t+1}. This sign assumes that the inverse basis and the excess return are negatively correlated. As before, Eq. (3) itself does not suggest the correlation between ΔIB_t and ΔXR_{t+1}. However, if the correlation between ΔIB_t and ΔSR_t is less than perfect, then we will have some correlation between ΔIB_t and ΔXR_{t+1}. ΔIB_t is the basis spread. ΔXR_{t+1} corresponds to the basis factor return.[8] Thus, if the correlation between ΔIB_t and ΔSR_t is less than perfect, we are justified to examine the prediction of the basis factor return by the basis spread. In the actual data, one finds that ΔSR_t and ΔXR_{t+1} are highly correlated, and these two variables are much more volatile than ΔIB_t. Thus, only a small fraction of the variation in ΔXR_{t+1} can be explained by the variation in ΔIB_t, and the goodness of fit of our predictive regression is not very high. Nonetheless, as we show later, ΔIB_t includes an

[8] ΔXR_{t+1} equals to the basis factor return only if the portfolio return is identical to the median return. In practice, the portfolio return is the average (not median) return, and the identity fails to hold. Nonetheless, these two quantities are similar to each other, conceptually and also numerically. See Appendix A1 for further discussion.

important piece of information regarding ΔXR_{t+1}, and its predictability is statistically and economically significant.

Implication of Recent Theoretical Models

Another justification for our predictive regression comes from recent theoretical models by Gorton et al. (2012) and by Yang (2013). A common idea in these two models is that the variation in the excess return and the basis is driven by the economy-wide production shock. Cross-sectional differences in the excess return and the basis are attributable to the cross-section of commodity characteristics. The implication is that the cross-sectional *spreads* in excess return and in basis are free from the influence of commodity characteristics and reveal the magnitude of the production shock.

In Yang's (2013) model, the basis is influenced by the production shock[9] and the investment rate. The production shock--denoted by Y_t--is a shock to the technology, and affects how much capital stock one unit of investment generates. The investment rate--denoted by $i_{j,t}$ for producer of commodity j--is different for different commodities, while the production shock Y_t is common. The product of the production shock and the investment rate, i.e. $Y_t i_{j,t}$, is the effective amount of investment. When the investment is high, the futures price (F_t) relative to the current spot price (S_t) declines, and the basis ($\mathrm{S_t}/F_t$) increases. At the same time, with high investment, the exposure to the future production shock is higher and the risk premium is also high. Thus, in this model, production shock drives both the basis and the risk premium. When we calculate the basis spread, the effect of $i_{j,t}$ is removed, and the spread is proportional to the production shock Y_t. Thus, some correlation between the basis spread and a measure of the production shock is to be expected. The basis factor has been shown to be one of

[9] Yang calls it "investment shock," which shows the increase in capital stock that one unit of investment generates.

the most important factors in the commodity futures market, and one may interpret it as a reflection of the economy-wide production shock. Thus, the basis factor return is likely to be correlated with the basis spread. Correlation will be stronger if the basis factor is the true common factor in the commodity futures market and thus a good proxy for the production shock.

In the model by Gorton et al. (2012), production and demand shocks determine the basis and the risk premium. The production shock--denoted by $\tilde{z}$--influences the supply of commodities and thus the next-period spot price. The demand shock--denoted by $\tilde{\varepsilon}$--influens the next-period spot price through the demand. These two shocks affect the basis and the risk premium via the variance of the next-period spot price and also the variance of hedgers' and speculators' profit. While the model by Gorton et al. considers one commodity only (and therefore does not allow statements regarding cross-sectional spreads), it is easy to extend the model to the case of multiple commodities, as we show in Appendix A2. All the ideas of the one commodity case extend to the case of multiple commodities. In particular, the basis of commodity j is proportional to the variance of the common production shock, $\tilde{z}$, and a small number of commodity-specific parameters. Then, the basis *spread* does not depend on commodity-specific parameters, and is proportional only to the variance of the common production shock. The common production shock also determines the factor premium. Therefore, once again, the factor return is likely to be correlated with the basis spread.

3. Predictive Regression Results

In this section, we discuss the predictive regression results. We first describe the data and the variables; then we present our key results followed by additional analyses and discussion.

Data

Our analysis is based on the commodity futures market data for the 40-year period between 1972 and 2011. We have collected futures price data from the Commodity Research Bureau database. We have selected 31 commodities with more than 10 years of price data.[10] See Appendix A3 for the list of the selected commodities.

We calculate the excess return and the inverse basis of each commodity in the following way. For the excess return calculation, we adopt the "nearest-maturity-contract formulation": for each day, we identify the nearest-maturity contract that has at least 5 calendar days remaining until the last trade date.[11] From the nearest-maturity contracts, we create a single price index for each commodity. We adjust the price index for any changes of contracts so that the rate of change in this index properly represents the investment return. The monthly return is calculated from this price index. We calculate the inverse basis out of the prices of the first- and the second-nearest-maturity contracts. The second-nearest-maturity contract is identified in the same way as the first-nearest-maturity contract, i.e. it has the nearest maturity among all the contracts that have at least 5 calendar days remaining until the last trade date and are not the first-nearest-maturity contract. The inverse basis is the ratio of the second-nearest-maturity-contract price to the first nearer-maturity-contract price. We make the adjustment for the interval between two maturity dates. Thus, the inverse basis is defined as $100\left[\left(F_{t,n_2}/F_{t,n_1}\right)^{1/(n_2-n_1)}-1\right]$, where n_1 and n_2 are the time-to-maturity (in months) of the nearer-maturity and the second nearest-maturity contracts, respectively, and F_{t,n_1} and F_{t,n_2} are the prices of these contracts.

[10] We carried out a preliminary analysis using the price data from Bloomberg, and obtained essentially the same results. It appears that our results are quite robust to changes in the time period as well as the list of commodities.

[11] That is, our "roll-over" date is 5 calendar dates prior to the expiration date. This corresponds to the traders' rollover strategy in practice. A rollover at the last possible moment faces greater uncertainty.

Note that we have decided not to use spot price in our calculation of the excess return and the inverse basis. Although we could have attempted to match futures contracts with the appropriate spot contracts, such matching is likely to introduce errors; moreover, for some commodities, such matching is impossible. Thus, we calculate the excess return and the basis out of the futures price data only. This is also the procedure adopted by many authors.[12] One consequence of not using spot price is that the "near identity" in Eq. (2) becomes less accurate. The deviation from the identity will be large if the term structure of the futures prices is far from being linear.[13] In any case, the motivation for the predictive regression does not require the relationship in Eq. (2) to be accurate.

The basis factor is created after sorting the commodities by the inverse basis.[14] We consider two variants: The "H1-H2" factor is comprised of the below-median and the above-median portfolios, and the "Q1-Q4" factor is made of the 1st quartile and the 4th quartile portfolios. In both cases, the low-inverse-basis (i.e., high-basis) portfolio constitutes the long side, whereas the high-inverse-basis (i.e., low-basis) portfolio constitutes the short side. The basis factor return is the difference between the long and the short portfolio returns. The portfolio is equally weighted, and is rebalanced every month.

[12] Gorton et al. (2012), De Roon and Szymanowska (2010), Yang (2013), Szymanowska et al. (2013), and Bakshi et al. (2014) are some of the studies adopting this procedure.

[13] If the spot price is used, the inverse basis can be calculated as $100\left[\left(F_{t,n_1}/S_t\right)^{1/n_1}-1\right]$, where F_{t,n_1} is the price of the nearest-maturity futures contract and n_1 is the time-to-maturity of this contract. The inverse basis that we calculate is $100\left[\left(F_{t,n_2}/F_{t,n_1}\right)^{1/(n_2-n_1)}-1\right]$, where F_{t,n_2} is the price of the nearest-maturity futures contract and n_2 is the time-to-maturity of this contract. These two quantities will be similar to each other if the three points, $(0,S_t)$, (n_1,F_{t,n_1}), and (n_2,F_{t,n_2}), are close to being on a line.

[14] The (ascending) sort by the inverse basis is of course identical to the descending sort by the basis. See footnote 2.

For the basis spread, we consider two variants as well[15]: The "75-25" spread is the inter-quartile spread, i.e., the difference between the 3rd and the 1st quartiles ("the 75th and the 25th percentiles"), and the "87.5-12.5" spread is the difference between the 7th and the 1st 8-quantiles ("the 87.5th and the 12.5th percentiles").[16]

Table 1 presents the summary statistics of the basis factor return and the basis spread. As expected, the average basis factor return is significantly positive. The monthly mean is 0.7% in the case of the "H1-H2" factor and 1.2% in the case of the "Q1-Q4" factor.[17] The basis spread is positive by construction and is about three times as large as the average basis factor return. The table also reports the auto-correlation in each variable. The basis spread variables are somewhat persistent, having the first-order auto-correlation of 0.63 and 0.48. The persistent right-hand-side variable in a regression may cause a problem; we argue later that this magnitude of the auto-correlation does not introduce a serious bias. Also interesting is the fact that the 12th order auto-correlation is not much different from the 3rd order auto-correlation. Considering that the variables are monthly, the non-zero 12th order auto-correlation suggests the existence of seasonality with annual frequency. To make sure that the seasonality is not driving our main results, we repeat a part of our analysis with seasonally adjusted variables, as we explain below.

[Insert Table 1 here.]

[15] Note that the basis spread is calculated out of the inverse basis. We use the inverse basis (rather than the basis) to be more consistent with Eq. (3).

[16] The "75-25" spread is comparable to the "H1-H2" factor in that the 3rd and the 1st quartiles are the mid-points of the below-median and the above-median ranges, respectively. Similarly, the "87.5-12.5" spread is comparable to the "Q1-Q4" factor because the 7th and the 1st 8-quantiles are the mid-points of the above-1st-quartile and the below-3rd-quartile ranges.

[17] These means are statistically significant. The T statistics for these means are 3.26 and 4.41, respectively.

Results

Our key predictive regression results are presented in Table 2. The basis factor return is regressed on the basis spread and the constant term. We have four variants of the basis spread (“75-25” and “87.5-12.5”, seasonally adjusted and non-adjusted), and also have the two variants of the basis factor return (“H2-H1” and “Q4-Q1”). Thus, there are total of 8 regressions. For each regression, we report OLS estimates and T statistics. We also report the T statistics after making the Newey-West (1987) adjustment for heteroskedasticity and serial-correlation. One can see from the table that the basis spread is highly significant. The “87.5-12.5” spread is significant at 1% in one regression and 5% in another, whereas the “75-25” spread is significant at 5% in one regression and 10% in another.

[Insert Table 2 here.]

Certain commodity prices exhibit seasonality; thus basis spread may contain seasonality as well. Indeed, the auto-correlation reported in Table 1 indicates such possibility: The auto-correlation of order 12 is quite high for the basis spread variables. When we test for the existence of the seasonality using the F test, the null hypothesis of no seasonality is rejected.[18] Given the possible presence of seasonality in the predictor variable and no apparent presence of seasonality in the dependent variable, one may suspect that removing seasonality in the predictor variable may improve the predictability. To see whether this is the case, we have repeated our predictive regression after seasonally adjusting

[18] The F test was done using the X11 procedure of Statistics Canada and the U.S. Bureau of the Census. See Dagum (1978). This test is meant to be only indicative. A more serious test requires a time-series model of the basis spread.

the predictor variable.[19] Table 2 includes the regression results with the seasonally adjusted predictors. We see minor improvement in the case of the basis spread "75-25", but overall, the results are comparable to the pre-adjustment patterns.

A related issue is the effect of persistent predictors on the estimated coefficient. In general, persistent predictors may introduce bias in the coefficient estimates. Fortunately, the magnitude of the potential bias can be determined from the sample size and the autocorrelation of the predictor. In our case, the sample size is relatively large and the autocorrelation is low; thus, the magnitude of the bias is likely to be very small.[20]

We have performed several robustness checks. For the consideration of space, we present one set of such analysis in Table 3. Here, we modify the specification by adding the lagged variables in the right-hand side of the regression equation. When the lagged predictor is added, the non-lagged predictor remains significant. When the lagged dependent variable is added, however, the original predictor becomes less significant. Nonetheless, the estimated coefficient remains above 0.5 indicating some resilience. Overall, while we do not obtain the statistical significance from all variations, the estimated coefficient is always positive and mostly comparable to the ones that we

[19] The seasonal adjustment was made again using the X11 procedure of Statistics Canada and the U.S. Bureau of the Census. See Dagum (1988).

[20] Stambaugh (1999), and also Nelson and Kim (1993), provide the following formula. Given the system of equations, $y_t = \alpha + \beta x_{t-1} + u_t$, $x_t = \mu + \phi x_{t-1} + v_t$, the bias in the OLS estimate of β is given as $\mathrm{E}(\hat{\beta} - \beta) = \sigma_{uv}/\sigma_v^2 \mathrm{E}(\hat{\phi} - \phi)$, where $\mathrm{E}(\hat{\phi} - \phi) = -(1 + 3\phi)/n + O(n^{-2})$. In the estimation of Table 2, sample size n is 480, and the estimate of the first-order serial correlation $\hat{\phi}$ is around 0.5 and 0.6 as shwon in Table 1. To get a rough estimate of the bias in $\hat{\phi}$, we can rewrite the bias formula in terms of $\hat{\phi}$: $\mathrm{E}(\hat{\phi} - \phi) \approx -3\hat{\phi}/(n-3) - 1/(n-3)$. Substituting $\hat{\phi} = 0.6$ and n=480 yields $\mathrm{E}(\hat{\phi} - \phi) \approx 0.006$. Thus, for most of the feasible values of σ_{uv}/σ_v^2, the bias in $\hat{\beta}$ is very small.

obtain in Table 2.

[Insert Table 3 here.]

Decomposition Interpretation of Regression Coefficients

Given the approximate identity in Eq. (3), we may interpret the coefficient on the basis spread as the share of the excess return component of the variation in the basis spread. Eq. (3) implies that

$$\mathrm{var}(\Delta IB_t) = \mathrm{cov}(\Delta IB_t, \Delta SR_t) + \mathrm{cov}(\Delta IB_t, \Delta XR_{t+1}) \qquad (4)$$

Dividing both sides by $\mathrm{var}(\Delta IB_t)$, we have

$$1 = \beta_{\Delta SR,\Delta IB} + \beta_{\Delta XR,\Delta IB} \qquad (5)$$

where $\beta_{\Delta SR,\Delta IB}$ is the coefficient in the (times-series) regression of ΔSR_t on ΔIB_t, and $\beta_{\Delta XR,\Delta IB}$ is the coefficient in the (time-series) regression of ΔXR_{t+1} on ΔIB_t. Given that $\beta_{\Delta SR,\Delta IB}$ and $\beta_{\Delta XR,\Delta IB}$ sum to one, we may interpret these quantities as representing the share of the spot return and the excess return components in the basis spread.[21]

Looking at Table 2 again, over 50% of the variation in the basis spread is attributable to the basis factor return. In regression (7) where the predictor is the basis spread “75-25” seasonally adjusted, over 80% of the variation in the basis spread is attributable to the basis factor return. Even in the lowest coefficient case, regression (2), the basis-factor-return share of the variation in the basis spread is not far below 50%. Considering this, we may say that the predictive regression result is economically meaningful.

[21] Fama (1984) and Fama and French (1987) have applied this decomposition idea to the analysis of FX and commodity futures markets. Cohen et al. (2003) also discuss this idea.

Sub-Period Results

For additional robustness check, we split the entire sample into two sub-periods, and then run regressions for each sub-period. Table 4 presents the sub-period regression results. For the first half of the sample, the coefficient estimate is only marginally significant, but its value is not very different from the full-sample estimate. For the latter half of the sample, although the coefficient estimate is not significant at the conventional level, its value is still positive.

[Insert Table 4 here.]

Why does the coefficient estimate become insignificant in the second half of the sample? One contributing factor is the reduced variance of the basis spread. (Recall that the coefficient estimate in OLS is inversely proportional to the variance of the right-hand-side variable.) The standard deviation of the basis spread in the first half of the sample is .0089, whereas it is .0060 in the second half of the sample.[22] What caused the decline in the standard deviation?[23] Further analysis indicates that the bases of individual commodities exhibit “the regression toward means” over time.[24] When we regress

[22] The reduction in the standard deviation can be visually confirmed from Figure 1.

[23] We have checked that the changed composition of the sample is not responsible for the reduced volatility. We have repeated the analysis excluding all the commodities that enter our database after year 1980, and we have obtained the same pattern. The standard deviation of the basis spread declines from 0.0088 in the first half of the sample to 0.0048 in the second half of the sample. So the changed composition of the sample is not a main factor here.

[24] This might be related to the financialization of the commodity market (Tang and Xiong (2012)), i.e., more investor participation might have reduced the basis spread volatility. Further analysis, however, indicates that the reduction in the volatility was the largest in 1990s rather than after 2000, so the timing does not coincide with the financialization wave.

individual commodities' average basis for the first half of the sample on the average basis for the second half of the sample, the slope coefficient is almost zero. Thus, those commodities with extreme basis in the first part of the sample have less extreme basis in the second half of the sample. And this contributes to the lack of significance in the second half of the sample.

4. Alternative Predictors

The basis spread is not the only potential predictor for the basis factor return. In fact, literature suggests a number of alternative predictors. The model of Gorton et al. (2012) implies the relevance of the business cycle since inventory tends to be counter-cyclical. Miffre and Rallis (2007), Shen et al. (2007), and Fuertes et al. (2010) discuss the relevance of the past return. Bessembinder (1992) and Gorton et al. (2012), among many others, have emphasized the role of spot price volatility in the futures price determination. Carter et al. (1983) and Bessembinder (1992) discuss hedging pressure.[25] Hong and Yogo (2012) show the predictability of the futures market interest. These variables can be thought of in the framework of the theory of storage. In the classical theory of storage, the inverse basis has three components: interest rate, storage cost, and convenience yield.[26] A major component of the convenience yield may be the benefit from avoiding *spot price volatility*, in which case the basis must be correlated with the volatility. The spot price volatility is influenced by the inventory

[25] The modern discussion of hedging pressure originates from, or at least is attributable to, the normal backwardation theory of Keynes (1930). Cootner (1960) is one of the early commentators who interpreted the normal backwardation idea as being driven by the hedging pressure.

[26] Let $F_{t,n}$ be the end-of-period-t price of the futures contract which expires at the end of period $\mathrm{t}+\mathrm{n}$. That is, n is the time-to-maturity of the contract (in years). S_t is the spot price at the end of period t. r is the deposit rate, u is the cost of storage, and c is the convenience yield. r, u, and c are expressed as continuously compounded annual rates. Then $F_{t,n} = S_t e^{(r+u-c)n}$

level through the option value of the inventory and the possibility of a stockout (as shown by Litzenberger and Rabinowitz (1995), for example). Further, the inventory level tends to be counter-cyclical, suggesting the role of *business cycle* in the determination of the basis. Or, as Hong and Yogo (2012) emphasized, when the risk absorption capacity is limited, the *market interest* may be a better indicator of the economic activity level. Also, the spread in *past return* is thought to reveal the stage of business cycle, as implied by Stivers and Sun (2010). The convenience yield may also reflect the imbalance between spot and futures markets. For example, buyers of commodities may prefer spot contracts to futures contracts because the former has the option value. This imbalance may create *quantity pressure* in the spot market, which then increases the spot price relative to the futures price. In this case, the basis reflects the quantity pressure.

Below, we first consider a possible role of the business cycle, and then examine the spreads in past return, spot price volatility, hedging pressure, and market interest. We also consider the interest rate variables that are known to predict the asset markets.[27]

Business Cycles

Figure 1 plots the basis spread and the basis factor return over the US business cycle. The contraction period (as determined by the National Bureau of Economic Research) is indicated by the shaded area. However, the counter-cyclicality of the basis spread is not very obvious from the graph. The same is true for the basis factor return. One can see that, during the recession periods, the basis factor return often exhibits an upward trend. In 5 out of 6 contraction periods in the past 40 years, the basis factor return recorded positive values. As will be shown later, however, counter-cyclicality is not visible in

[27] See, for example, Hong and Yogo (2012) for how these variables are used in the context of the asset market prediction.

the regression analysis when we use the industrial production as an indicator of business cycle.

[Insert Figure 1 here.]

We note that both the basis spread and the basis factor return look very volatile in the first few years of the sample period. This is partly due to the small cross-sectional size (N) of the sample for these early years. We have verified that our key results (of this section as well as of Sections 3 and 4) are not mainly driven by this period; We have repeated our analysis excluding this period, and we have obtained mostly comparable results.

Alternative Predictors

We consider the total of 9 predictor variables: (i) the short rate, (ii) the TED premium, (iii) the term premium, (iv) the corporate yield premium, (v) industrial production, (vi) the past return spread, (vii) the implied volatility spread, (viii) the hedging pressure spread, (ix) the market interest spread. The first four variables are based on interest rates and bond yields. The short rate is the yield on the 3-month treasury bill. The TED premium is the difference between the short rate and the 3-month LIBOR. The term premium is the yield difference between the 3-month treasury bill and the 10-year treasury bond. The corporate yield premium is the Moody's Aaa corporate bond yield over the short rate. The industrial production variable is included to capture the effect of the business cycle. It is calculated as the rate of change in the seasonally adjusted index. The remaining four variables are the inter-quartile spread on commodity characteristics. The monthly average of the past-12-month excess return is our measure of the "past return." Our measure of spot price volatility is the implied volatility from the futures options. We have obtained the implied volatility of each commodity from the Commodity Research Bureau database. The hedging pressure is calculated as hedgers' net long position relative to all open interest. The underlying data come from the Commitment of Traders reports.[28] The market interest is the rate of change in the dollar value of all open interest. For each of the past return, the implied volatility, the hedging pressure, and the market interest, we calculate the inter-quartile spread. The past return spread can be calculated for the full-sample period (1972-2011). The three other spread variables are calculated for the second half of the sample period (1991-2011) due to the lack of data availability.

Table 5 reports the univariate statistics, and Table 6 shows the correlation among the predictor variables. The largest correlation is between the basis spread and the past return spread. This correlation is 0.33; all other correlations are smaller. The correlation table foretells the result of the predictive regression to be discussed shortly. Out of the four spread variables, the past return spread has the highest predictive power for the basis factor return.

[Insert Table 5 here.]

[Insert Table 6 here.]

Table 7 reports the predictive regression results based on alternative predictors. Due to limited data availability, the last four regressions involving these variables are based on the second-half of the sample (1991-2011). It turns out that the short rate and the TED premium are significant predictors of the basis factor return. Among the spread variables, the past return spread is the only (marginally) significant predictor. As for the basis spread, it remains significant when other predictors are added to the equation as long as the full sample is used. In the second half of the sample, no predictor is

[28] Following the convention, we identify ‘commercial traders’ as hedgers. By ignoring the positions of ‘non-reportable’ we are implicitly assuming that small traders are speculators. The Commitment of Traders reports are released on a weekly basis, with more than one week’s delay. The end-of-month hedging pressure is based on the last weekly report of the month, without considering the release gap.

significant.[29] Even in the second half of the sample, however, the estimated coefficients remain quite stable.

[Insert Table 7 here.]

5. Conclusion

We apply the spread-predicts-the-factor-return idea to the commodity futures market. We find that the inter-quartile spread in the commodity-futures basis predicts the basis factor return. From the underlying identity relation, Eq. (2), we can infer that over 50% of the time variation in the basis spread is attributable to the variation in the basis factor return. The predictability is very strong when the full sample (1972 ~ 2011) is considered. It is less strong in the second half of the sample (1991-2011) possibly due to the tendency of the basis to “regress toward means” over time. Introducing other predictors does not change the main results. Our finding supplements the existing evidence regarding the relationship between the basis and the excess return of commodity futures. Also, it is consistent with the theoretical models of commodity futures that Yang (2013) and Gorton et al. (2012) proposed. In these models, the risk premium and the basis are driven by the economy-wide production shock, and the basis spread is a good indicator of the magnitude of the shock.

Appendix

A1. Derivation and Interpretation of Eq. (3)

We use the following notation and terminology. We denote the k-th q-quantile of X by $x^{(\mathrm{q,k})}$. It is

[29] See the discussion following Table 4.

obtained from $\Pr\left(X \leq x^{(q,k)}\right) = k/q$. When $q = 4$, the q-quantiles are called quartiles. The inter-quartile spread of X is the difference between the 3rd and the 1st quartiles, i.e., $\Delta\mathrm{X} = x^{(4,3)} - x^{(4,1)}$.

Two variables X and Y have a monotonic relationship if whenever x_i is higher than x_j, y_i is higher than y_j. If X and Y have a monotonic relationship, then the k-th q-quantiles of X and Y, $x^{(q,k)}$ and $y^{(q,k)}$, obtain simultaneously. If X and $-\mathrm{Y}$ have a monotonic relationship, then $x^{(q,k)}$ and $y^{(q,q-k)}$ obtain at the same time. (Note that the k-th q-quantile of $-\mathrm{Y}$ equals the minus of the (q − k)-th q-quantile of Y.) Now, suppose that *IB* and *SR* have a monotonic relationship. Suppose also that *IB* and $-XR$ have a monotonic relationship. Then the k-th q-quantiles of *IB* and *SR,* and the (q − k)-th q-quantile of *XR* obtain at the same time. This fact, together with Eq. (2), implies

$$IB^{(\mathrm{q,k})} \approx SR^{(q,k)} - XR^{(q,q-k)}$$

Substituting (4,3) and (4,1) for (q,k) in turn, and then by taking the difference,

$$IB^{(4,3)} - IB^{(4,1)} \approx \left(SR^{(4,3)} - SR^{(4,1)}\right) - \left(XR^{(4,1)} - XR^{(4,3)}\right)$$

Noting that the inter-quartile spread of X is $\Delta\mathrm{X} = x^{(4,3)} - x^{(4,1)}$, we obtain Eq. (4).

The excess return spread $XR^{(4,3)} - XR^{(4,1)}$ is nearly identical to the basis factor return conceptually. The numerical values of these two quantities are likely to be similar to each other for the following reason. Let $XR_{\mathrm{t+1}}^{(2,1)}$ be the median excess return. Then

$$XR_{\mathrm{t+1}}^{(4,3)} = \mathrm{med}\left(XR_{i,t+1} | XR_{i,t+1} \geq XR_{t+1}^{(2,1)}\right)$$

$$XR_{\mathrm{t+1}}^{(4,1)} = \mathrm{med}\left(XR_{i,t+1} | XR_{i,t+1} \leq XR_{t+1}^{(2,1)}\right)$$

If the excess return and the inverse basis have a monotonic relationship, then the condition $XR_{i,t+1} \geq XR_{t+1}^{(2,1)}$ can be replaced by $IB_t \leq IB_t^{(2,1)}$; also the condition $XR_{i,t+1} \leq XR_{t+1}^{(2,1)}$ can be replaced by $IB_t \geq IB_t^{(2,1)}$. Thus,

$$XR_{\mathrm{t+1}}^{(4,3)} = \mathrm{med}\left(XR_{i,t+1} | IB_t \leq IB_t^{(2,1)}\right)$$

$$XR_{\mathrm{t+1}}^{(4,1)} = \mathrm{med}\left(XR_{i,t+1} | IB_t \geq IB_t^{(2,1)}\right)$$

That is, $XR_{\mathrm{t+1}}^{(4,3)}$ is the median return of the low-inverse-basis portfolio, and $XR_{\mathrm{t+1}}^{(4,1)}$ is the median return of the high-inverse-basis portfolio. Then, $XR^{(4,3)} - XR^{(4,1)}$ is very close to the basis factor return that we calculate in our empirical analysis, as long as the median return (used in the spread calculation) is not very different from the average return (used in the factor return calculation).

A2. Extension of the Model by Gorton et al. (2012)

The model by Gorton et al. (2012) describes the risk premium and the basis of one commodity. We extend the model to the case of multiple commodities. Our goal is to show that the basis spread is proportional to the variance of the production shock.

There exist a representative hedger (producer) and a representative speculator, as in the original model. Unlike in the original model, there exist K commodities, indexed by $j = 1, \cdots, K$. It may be more natural to assume one representative hedger for each commodity; this does not change the conclusion. So we choose the assumption that leads to the simplest formula.

Hedgers

There are two periods. In the first period, the hedger sets aside some of the initial inventory for the next period; i.e. x out of V is carried to the next period, where both x and V are K-vectors. The rest, $V - x$, is sold to the spot market. $S(\quad)$ is the inverse demand function (which is also K-dimensional). Thus, the hedger's first period profit is

$$\Pi_0 = S(V - \bar{x})'(V - x)$$

where $\bar{x}$ is the carryover amount averaged across all hedgers. In equilibrium, $\bar{x} = x$. The hedgers take a short position in the futures market. The position size is N, and futures price F. Both N and F are K-vectors. Between the first and the second period, two things happen. First, the inventory depreciates at the rate of δ so that only $(I-\delta)x$ is passed to the second period. δ is a diagonal matrix where j-th diagonal element indicates the depreciation rate of the j-th commodity. Second, there is production shock $\tilde{z}$ and demand shock $\tilde{\varepsilon}$. Then the hedger's second-period profit becomes

$$\Pi_1 = S(\tilde{z}+(I-\delta)\bar{x}-\tilde{\varepsilon})'(\tilde{z}+(I-\delta)x-N)+F'N$$

The interest rate is assumed to be zero. So the hedger's problem is to choose x and N so that

$$\max \Pi_0 + E(\tilde{\Pi}_1) - \frac{\alpha}{2}\mathrm{var}(\tilde{\Pi}_1) \quad \mathrm{s.t.} \quad x \geq 0$$

where α is the risk aversion parameter. The first order condition, which is derived through the straightforward application of the Lagrangian method, includes two sets of equations. In the first set, the depreciation rate (i.e. storage cost) has a prominent role:

$$S_0 = (I-\delta)F + \lambda$$

where $S_0 = S(V-\bar{x})$ and λ is the vector of Lagrange multipliers. Elements of λ are nonnegative; they are zero if the constraints are not binding, i.e. if $x_k > 0$ then $\lambda_k = 0$. The second part of the first order condition includes the formula for the risk premium:

$$E(\tilde{S}_1) - F = \alpha\{\mathrm{var}(\tilde{S}_1)[(I-\delta)x-N] + \mathrm{cov}(\tilde{S}_1{}'\tilde{z}, \tilde{S}_1)\}$$

where $\tilde{S}_1 = S(\tilde{z}+(I-\delta)\bar{x}-\tilde{\varepsilon})$.

Speculators

The speculator creates a portfolio of futures position. Given position N, the speculator's profit is

$$\tilde{\mathrm{W}} = (\tilde{S}_1 - F)N$$

Thus, the speculator's problem is to choose N so that

$$\max E(\tilde{W}) - \frac{\beta}{2}\mathrm{var}(\tilde{W})$$

where β is the risk aversion parameter. The first order condition to this problem is

$$N = \frac{1}{\beta}\left[\mathrm{var}\left(\tilde{S}_1\right)\right]^{-1}\left[E\left(\tilde{S}_1\right) - F\right]$$

Equilibrium

Combining the speculator’s first-order condition with the hedger’s first-order condition, the equilibrium condition can be described by the following formula:

$$S_0 = (I - \delta)F + \lambda$$

$$E\left(\tilde{S}_1\right) - F = \frac{\alpha\beta}{\alpha + \beta}\left\{\mathrm{var}\left(\tilde{S}_1\right)(I - \delta)x + \mathrm{cov}\left(\tilde{S}_1{}'\tilde{z}, \tilde{S}_1\right)\right\}$$

$$\bar{x} = x$$

Basis Spread

To illustrate our point, we now introduce a few simplifying assumptions. First, we set $\tilde{\varepsilon} = 0$, to focus on the production shock. Second, we assume that the third moment of $\tilde{z}$ is zero, i.e. $\mathrm{E}\left(\tilde{z}_k^2\tilde{z}_l\right) = 0$ for any k and l. Third, we assume a linear inverse demand such that $S_k(q_k) = c_k - d_k q_k$ or in vector form $S(q) = c - dq$ where d is a diagonal matrix. Then the equilibrium condition becomes

$$c - d(V - x) = (I - \delta)F + \lambda$$

$$c - d(I - \delta)x - F = \frac{\alpha\beta}{\alpha + \beta}\left[d\mathrm{var}(\tilde{z})c\right]$$

$$\bar{x} = x$$

For commodity k, the equilibrium condition is:

$$c_k - d_k(V_k - x_k) = (1 - \delta_k)F_k + \lambda_k$$

$$c_k - d_k(1 - \delta_k)x_k - F_k = \frac{\alpha\beta}{\alpha + \beta}d_k[\mathrm{var}(\tilde{z})]_{k\cdot}c$$

Define the basis as $1 + b_k = S_{0,k}/F_k = [c_k - d_k(V_k - x_k)]/F_k$, the equilibrium condition becomes

$$1 + b_k \geq 1 - \delta_k \quad (\text{"=" if } x_k > 0)$$

$$\frac{1}{1+b_k}=\frac{c_k-d_k(1-\delta_k)x_k-\dfrac{\alpha\beta}{\alpha+\beta}d_k[\mathrm{var}(\tilde{z})]_{k\cdot}c}{c_k-d_k(V_k-x_k)}$$

We first check the solution where $x_k > 0$. Then

$$1+b_k=1-\delta_k$$

$$\frac{1}{1+b_k}=\frac{c_k-d_k(1-\delta_k)x_k-\dfrac{\alpha\beta}{\alpha+\beta}d_k[\mathrm{var}(\tilde{z})]_{k\cdot}c}{c_k-d_k(V_k-x_k)}$$

This solution is less interesting for us as the basis spread is simply the spread in δ_k. The other solution is obtained if $x_k = 0$. If $x_k = 0$, then

$$\frac{1}{1+b_k}=\frac{c_k-\dfrac{\alpha\beta}{\alpha+\beta}d_k[\mathrm{var}(\tilde{z})]_{k\cdot}c}{c_k-d_kV_k}$$

A requirement is that $1+b_k \geq 1-\delta_k$.

Now suppose that $\mathrm{var}(\tilde{z})=\sigma^2 I$. (As can be seen in the next formula, this assumption helps to clarify the relationship between the basis spread and the variance of the production shock. A more general covariance matrix structure can be adopted with somewhat more cumbersome expression.) For the spread calculation, suppose that the first and the third quartiles are obtained from commodity k and k'. Note that we calculate the basis spread out of the inverse basis $1/(1+b_k)$. Then the basis spread (i.e. the spread in the inverse basis) is

$$\left(\frac{c_k}{c_k-d_kV_k}-\frac{c_{k\prime}}{c_{k\prime}-d_{k\prime}V_{k\prime}}\right)-\frac{\alpha\beta}{\alpha+\beta}\left(\frac{d_kc}{c_k-d_kV_k}-\frac{d_{k'}c}{c_{k'}-d_{k'}V_{k'}}\right)\sigma^2$$

This formula shows that the basis spread is proportional to the variance of the production shock σ^2.

A3. Data Details

Our commodity futures price data are from the Commodity Research Bureau database. The data have been used by many, including Gorton and Rouwenhorst (2006), Gorton et al. (2012), Shen et al. (2007), Yang (2013), and Szymanowska et al. (2013). We have excluded the commodities whose price series start after 2002. We have also excluded the metal commodities that are traded on the London Metal Exchange. The CRB data on these commodities are limited (i.e. there are no contract-by-contract data for these commodities.) After the selection, we have 31 commodities listed in Table A1. Table A1 shows the univariate statistics on monthly returns and end-of-the-month basis for each commodity in our sample. One can see that the average monthly returns are mostly positive, between 0% and 1%. The average inverse basis is mostly positive as well, suggesting that the commodity markets are more likely to exhibit an upward-sloping term structure.

[Insert Table A1 here.]

Table A2 presents the univariate statistics on three other characteristics that we examine in Section 5. The implied volatility data are also from the CRB database. The hedging pressure variable has been constructed from the CFTC Commitment of Traders reports. The implied volatility and the hedging pressure variables are available for a shorter period of time. For butter and propane, the implied volatility data are not available.

[Insert Table A2 here.]

References

Asness, Clifford S., Jacques A. Friedman, Robert J. Krail, and John M. Liew (2000), "Style Timing: Value versus Growth," *Journal of Portfolio Management*, 26(3), 50-60.

Bakshi, Gurdip, Xiaohui Gao, and Alberto Rossi (2014), “A Better Specified Asset Pricing Model to Explain the Cross-Section and Time-Series of Commodity Returns,” working paper.

Bessembinder, Hendrik (1992), “Systematic Risk, Hedging Pressure, and Risk Premiums in Futures Markets,” *Review of Financial Studies*, 5(4), 637~667.

Carter, Colin A., Gordon C. Rausser, and Andrew Schmitz (1983), “Efficient Asset Portfolios and the Theory of Normal Backwardation,” *Journal of Political Economy*, 91(2), 319-331.

Cohen, Randolph B., Christopher Polk, and Tuomo Vuolteenaho (2003), "The Value Spread," *Journal of Finance*, 58(2), 609-641.

Cootner, Paul H. (1960), "Returns to Speculators: Telser versus Keynes," *Journal of Political Economy*, 68(4), 396~404.

Dagum, Estela Bee (1978), “Modeling, Forecasting and Seasonally Adjusting Economic Time Series with the X-11 ARIMA Method,” *Journal of the Royal Statistical Society Series D*, 27(3), 203-216.

-------- (1988), *The X-11-ARIMA/88 Seasonal Adjustment Method: Foundations and User’s Manual*, Statistics Canada.

De Roon, Frans, and Marta Szymanowska (2010), “The Cross-Section of Commodity Futures Returns,” working paper.

Fama, Eugene F. (1984), “Forward and spot exchange rates,” *Journal of Monetary Economics*, 14, 319-338.

Fama, Eugene F., and Kenneth R. French (1987), “Commodity Futures Prices: Some Evidence on Forecast Power, Premiums, and the Theory of Storage,” *Journal of Business*, 60(1), 1987, 55~73.

Fuertes, Ana-Maria, Joelle Miffre, and Georgios Rallis (2010), “Tactical Allocation in Commodity Futures Markets: Combining Momentum and Term Structure Signals,” working paper.

Gorton, Gary B., Fumio Hayashi, and K. Geert Rouwenhorst (2012), "The Fundamentals of Commodity Futures Returns," *Review of Finance*, 17, 35-105.

Gorton, Gary B., and K. Geert Rouwenhorst (2006), “Facts and Fantasies about Commodity Futures,” *Financial Analysts Journal*, 62, 47-68.

Hansen, Lars Peter, and Robert J. Hodrick (1980), "Forward Exchange Rate as Optimal Predictors of Future Spot Rates: An Econometric Analysis," *Journal of Political Economy*, 88(5), 829-853.

Keynes, John Maynard (1930), *A Treatise on Money* Vol. 2, London: Macmillan. (Reprinted as The *Collected Writings of John Maynard Keynes* Vol. 5 by Cambridge University Press in 1978.)

Litzenberger, Robert H., and Nir Rabinowitz (1995), "Backwardation in Oil Futures Markets: Theory and Empirical Evidence," *Journal of Finance*, 50(5), 1517-1545.

Miffre, Joëlle, and Georgios Rallis (2007), "Momentum Strategies in Commodities Futures Markets," *Journal of Banking and Finance*, 31, 1863-1886.

Newey, Whitney K., and Kenneth D. West (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Constant Covariance Matrix," *Econometrica*, 55(3), 703-708.

Shen, Qian, Andrew C. Szarkmary, and Subhash C. Sharma (2007), "An Examination of Momentum Strategies in Commodities Futures Markets," *Journal of Futures Markets*, 27(3), 227-256.

Stambaugh, Robert F. (1999), "Predictive Regressions," *Journal of Financial Economics*, 54, 375-421.

Stivers, Chris, and Licheng Sun (2010), "Cross-Sectional Return Dispersion and Time-Variation in Value and Momentum Premia," *Journal of Financial and Quantitative Analysis*, 45(4), 987-1014.

Szymanowska, Marta, Frans de Roon, Theo Nijman, and Rob Van Den Goorbergh (2013), "An Anatomy of Commodity Futures Risk Premia," working paper.

Tang, Ke, and Wei Xiong (2012), "Index Investment and the Financialization of Commodities," *Financial Analysts Journal*, 68(6), 54-74.

Yang, Fan (2013), "Investment Shocks and the Commodity Basis Spread," *Journal of Financial Economics*, 110, 164-184.

Tables and Figures

Table 1. Basis Factor Return and Basis Spread

We calculate the basis factor return as the excess return to the long-short portfolio of high-basis commodity futures in the long side and low-basis commodity futures in the short side. For the “H1-H2” factor, the commodities in the top half of the basis ranking are included in the long side, whereas the commodities in the bottom half are included in the short side. For the “Q1-Q4” factor, the commodities in the top and bottom quartiles are selected. The portfolio is equally weighted, and is rebalanced every month. The basis spread is a measure of the cross-sectional dispersion in the basis. The “75-25” spread is the difference between the 3rd and the 1st quartiles (“the 75th and the 25th percentiles”) of the inverse basis. The “87.5-12.5” spread is calculated out of the 7th and the 1st 8-quantiles (“the 87.5th and the 12.5th percentiles”). See the text for the description of how individual commodities’ excess return and inverse basis are calculated.

	# obs (start date)	Mean	SD	Auto-correlation			
				lag = 1	2	3	12
Basis factor return							
"H2-H1"	480 (Jan 1972)	0.007	0.040	0.093	0.063	0.026	-0.032
"Q4-Q1"	480 (Jan 1972)	0.012	0.058	0.128	0.073	-0.007	-0.045
Basis spread							
"75-25"	480 (Jan 1972)	0.018	0.008	0.626	0.516	0.397	0.367
"87.5-12.5"	480 (Jan 1972)	0.037	0.012	0.478	0.329	0.205	0.206

Table 2. Predictive Regression

Each column of the table is a separate regression. The dependent variables are shown in the top row, and the explanatory variables are shown in the left-most column. All the regressions are based on monthly data between 1972 and 2011. For each explanatory variable, the table shows the OLS estimate followed by the unadjusted t statistics inside the square brackets and the Newey-West adjusted t statistics inside the curly brackets. The significance at 10%, 5%, and 1% is indicated by *, **, and ***, respectively. See the notes for Table 1 for the description of variables. S.a. indicates seasonal adjustment.

	Dependent var:							
	Basis factor return "H2-H1"				Basis factor return "Q4-Q1"			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	-0.005	-0.011	-0.007	-0.012	-0.002	-0.008	-0.003	-0.007
	[-1.11]	[-1.85]	[-1.42]	[-1.91]	[-.30]	[-.93]	[-.45]	[-.72]
		*		*				
	{-.93}	{-1.66}	{-1.17}	{-1.70}	{-.26}	{-.85}	{-.36}	{-.61}
		*		*				
Basis spread	0.658				0.765			
"75-25"	[2.72]				[2.19]			
	***				**			
	{2.09}				{1.79}			
	**				*			
Basis spread		0.474				0.531		
"87.5-12.5"		[3.09]				[2.39]		
		***				**		
		{2.65}				{2.14}		
		***				**		
Basis spread			0.755				0.829	
"75-25" s.a.			[2.97]				[2.25]	
			***				**	
			{2.27}				{1.68}	
			**				*	
Basis spread				0.505				0.495
"87.5-12.5" s.a.				[3.06]				[2.06]
				***				**
				{2.59}				{1.69}
				***				*
T	480	480	480	480	480	480	480	480
R sq.	0.0153	0.0196	0.0181	0.0192	0.0099	0.0118	0.0105	0.0088

Table 3. Predictive Regression: Alternative Specifications

Each column of the table is a separate regression. The dependent variable is the basis factor return “H1-H2,” and the explanatory variables are shown in the left-most column. All the regressions are based on monthly data between 1972 and 2011. The table shows the OLS estimate followed by the unadjusted t statistics inside the square brackets and the Newey-West adjusted t statistics inside the curly brackets. The significance at 10%, 5%, and 1% is indicated by *, **, and ***, respectively. See the notes for Table 1 for the description of variables.

	(1)	(2)
Intercept	-0.0045	-0.0053
	[-.9456]	[-1.0223]
	{-.8377}	{-.8993}
Basis spread	0.5884	0.5130
"75-25"	[2.4016] **	[1.6450] *
	{1.9839} **	{1.5083}
Basis spread	0.0737	0.1212
"75-25," lagged	[1.6059]	[.3914]
	{1.5757}	{.3859}
Basis factor		0.0734
"H1-H2," lagged		[1.5974]
		{1.5741}
# obs.	480	480
R sq.	0.0206	0.0209

Table 4. Predictive Regression: Sub-periods

Each column of the table is a separate regression. The dependent variable is the basis factor return "H1-H2," and the explanatory variables are shown in the left-most column. The table shows the OLS estimate followed by the unadjusted t statistics inside the square brackets and the Newey-West adjusted t statistics inside the curly brackets. The significance at 10%, 5%, and 1% is indicated by *, **, and ***, respectively. See the notes for Table 1 for the description of variables.

	Sample period: Jan 1972-Dec 1990	Jan 1991-Dec 2011
	(1)	(2)
Intercept	-0.0059	-0.0034
	[-.8451]	[-.5078]
	{-.7248}	{-.5467}
Basis spread "75-25"	0.7104	0.5342
	[2.1409] **	[1.4171]
	{1.6925} *	{1.5193}
# obs.	228	252
R sq.	0.0199	0.0080

Table 5. Alternative Predictors

The short rate is the yield on the 3-month treasury bill. The TED premium is the difference between the short rate and the 3-month LIBOR. The term premium is the yield difference between the 3-month treasury bill and the 10-year treasury bond. The corporate yield premium is the Moody's Aaa corporate bond yield over the short rate. Industrial production is the rate of change in the seasonally adjusted index. Past return refers to the average of the monthly excess returns of the previous 12 months. Implied volatility is obtained from the nearest-maturity option prices. Hedging pressure is calculated as hedgers' net long position relative to all open interest, as presented in the Commitment of Traders reports. Market interest is the rate of change in the dollar value of all open interest. The spread variables are the differences between the 3rd and the 1st quartiles ("the 75th and the 25th percentiles) of the underlying variables.

	# obs (start date)	Mean	SD	Auto correlation			
				lag = 1	2	3	12
Short rate	480 (Jan 1972)	0.053	0.032	0.962	0.933	0.914	0.756
TED premium	480 (Jan 1972)	0.011	0.011	0.750	0.676	0.637	0.475
Term premium	480 (Jan 1972)	0.018	0.014	0.847	0.767	0.717	0.324
Corporate yield premium	252 (Jan 1991)	0.033	0.015	0.940	0.917	0.902	0.525
Industrial production	480 (Jan 1972)	0.002	0.008	0.371	0.327	0.307	-0.021
Past return spread	480 (Jan 1972)	0.034	0.013	0.812	0.682	0.591	0.016
Implied volatility spread	252 (Jan 1991)	0.140	0.038	0.526	0.430	0.314	0.092
Hedging pressure spread	252 (Jan 1991)	0.100	0.025	0.635	0.445	0.392	0.276
Market interest spread	252 (Jan 1991)	0.197	0.054	0.204	0.168	0.267	0.140

Table 6. Correlation among Basis Spread and Other Predictors

Correlations involving corporate yield spread, implied volatility spread, hedging pressure spread, and market interest spread are calculated out of monthly data between 1991 and 2011. All other correlations are calculated out of monthly data between 1972 and 2011. See the notes on Tables 1 and 5 for the description of each variable.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Basis spread "75-25"	(1)	1.00	0.08	0.09	-0.14	-0.02	-0.09	0.33	0.19	-0.10	0.20
Short rate	(2)		1.00	0.47	-0.49	-0.80	-0.02	0.18	-0.07	0.07	0.09
TED premium	(3)			1.00	-0.08	0.09	-0.17	0.28	0.07	-0.10	-0.17
Term premium	(4)				1.00	0.95	0.07	-0.28	0.00	0.09	0.06
Corporate yield premium	(5)					1.00	-0.10	-0.12	0.13	-0.02	0.01
Industrial production	(6)						1.00	-0.14	-0.16	0.18	0.03
Past return spread	(7)							1.00	0.16	-0.12	-0.06
Implied volatility spread	(8)								1.00	-0.17	-0.04
Hedging pressure spread	(9)									1.00	0.20
Market interest spread	(10)										1.00

Table 7. Predictive Regression: Alternative Predictors

Each column of the table is a separate regression. The dependent variable is the basis factor return “H1-H2,” and the explanatory variables are shown in the left-most column. The table shows the OLS estimate followed by the Newey-West adjusted t statistics inside the curly brackets. The significance at 10%, 5%, and 1% is indicated by *, **, and ***, respectively. See the notes on Tables 1 and 5 for the description of each variable.

	Sample period:							
	Jan 1972-Dec 2011				Jan 1991-Dec 2011			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Basis spread “75-25”	0.65	0.71	0.65	0.49	0.59	0.57	0.61	0.57
	{2.10}	{2.23}	{2.12}	{1.66}	{1.40}	{1.36}	{1.39}	{1.33}
	**	**	**	*				
Short rate	-0.13	-0.02	-0.13	-0.14	-0.06	-0.06	-0.07	-0.07
	{-1.93}	{-.34}	{-1.93}	{-2.03}	{-.30}	{-.30}	{-.33}	{-.34}
	*		*	**				
TED premium	0.43		0.43	0.35				
	{2.36}		{2.28}	{1.86}				
	**		**	*				
Term premium		0.17						
		{1.05}						
Corporate yield premium					0.12	0.11	0.11	0.11
					{.48}	{.47}	{.44}	{.44}
Industrial production,			0.02					
			{.07}					
Past return spread				0.31				
				{1.86}				
				*				
Implied volatility spread						0.02		
						{.27}		
Hedging pressure spread							0.02	
							{.28}	
Market interest spread								0.02
								{.44}
T	480	480	480	480	252	252	252	252
R sq.	0.029	0.020	0.029	0.036	0.014	0.014	0.014	0.015

Table A1. Excess Return and Basis by Commodity

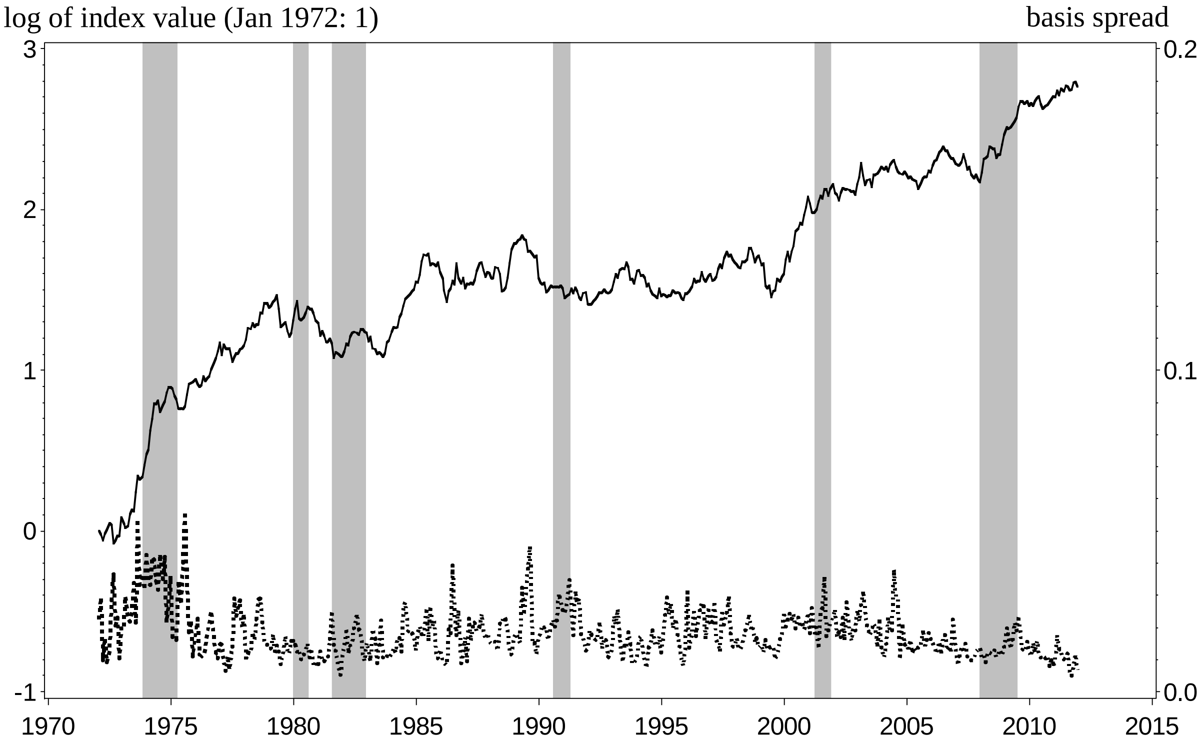
Commodity	Start date	End date	# obs.	Excess return		Inverse basis	
				Mean	SD	Mean	SD
Agricultural commodities							
Butter	OCT1996	NOV2010	170	0.0041	0.0954	0.0054	0.0213
Cocoa	JAN1972	DEC2011	480	0.0079	0.0965	0.0014	0.0152
Coffee	SEP1972	DEC2011	472	0.0083	0.1108	0.0020	0.0177
Corn	JAN1972	DEC2011	480	0.0018	0.0786	0.0077	0.0149
Feeder cattle	JAN1972	DEC2011	480	0.0056	0.0498	-0.0003	0.0145
Lean hogs	JAN1972	DEC2011	480	0.0095	0.0831	0.0082	0.0460
Live cattle	JAN1972	DEC2011	480	0.0077	0.0549	0.0005	0.0199
Milk	FEB1996	DEC2011	189	0.0031	0.0410	-0.0014	0.0623
Oats	JAN1972	DEC2011	480	0.0067	0.1018	0.0070	0.0209
Orange juice	JAN1972	DEC2011	480	0.0077	0.0928	0.0030	0.0157
Pork bellies	JAN1972	JUN2011	474	0.0074	0.1152	-0.0039	0.0193
Rough rice	SEP1986	DEC2011	304	0.0014	0.0871	0.0105	0.0183
Soybean meal	JAN1972	DEC2011	480	0.0092	0.1006	-0.0005	0.0220
Soybean oil	JAN1972	DEC2011	480	0.0086	0.0966	0.0011	0.0181
Soybeans	JAN1972	DEC2011	480	0.0069	0.0854	0.0016	0.0163
Sugar	JAN1972	DEC2011	480	0.0032	0.1260	0.0030	0.0208
Wheat	JAN1972	DEC2011	480	0.0014	0.0817	0.0055	0.0168
Non-agricultural commodities							
Blendstock gasoline	APR1985	DEC2011	321	0.0177	0.1108	-0.0056	0.0316
Coal	AUG2001	DEC2011	125	0.0027	0.0860	0.0067	0.0217
Copper	JAN1972	DEC2011	480	0.0105	0.0843	-0.0002	0.0128
Cotton	JAN1972	DEC2011	480	0.0046	0.0788	0.0023	0.0212
Crude oil	APR1983	DEC2011	345	0.0117	0.1004	-0.0019	0.0197
Gold	FEB1975	DEC2011	443	0.0036	0.0565	0.0047	0.0031
Heating oil	FEB1979	DEC2011	395	0.0118	0.1009	-0.0001	0.0240
Lumber	JAN1972	DEC2011	480	-0.0031	0.0912	0.0097	0.0269
Natural gas	MAY1990	DEC2011	260	-0.0048	0.1567	0.0204	0.0664
Palladium	FEB1977	DEC2011	419	0.0105	0.1052	0.0017	0.0057
Platinum	JAN1972	DEC2011	480	0.0067	0.0809	0.0025	0.0048
Propane	OCT1987	AUG2009	263	0.0211	0.1727	-0.0063	0.0282
Silver	JAN1972	DEC2011	480	0.0078	0.1007	0.0052	0.0062
Unleaded gas	APR1985	DEC2006	261	0.0189	0.1122	-0.0075	0.0320

Table A2. Past Return, Implied Volatility, and Hedging Pressure by Commodity

Commodity	Past return			Implied Volatility			Hedging Pressure		
	(Jan 1972 ~ Dec 2011)			(Jan 1991 ~ Dec 2011)			(Jan 1991 ~ Dec 2011)		
	# obs.	Mean	SD	# obs.	Mean	SD	# obs.	Mean	SD
Agricultural commodities									
Butter	159	0.0040	0.0347				41	-0.0297	0.0744
Cocoa	480	0.0084	0.0317	252	0.3233	0.0768	252	-0.0334	0.0461
Coffee	460	0.0084	0.0351	252	0.3930	0.1109	252	-0.0510	0.0502
Corn	480	0.0015	0.0237	252	0.2634	0.0862	252	-0.0036	0.0535
Feeder cattle	469	0.0054	0.0173	252	0.1204	0.0412	252	0.0385	0.0595
Lean hogs	480	0.0099	0.0264	252	0.2317	0.0922	252	0.0035	0.0497
Live cattle	480	0.0079	0.0183	252	0.1519	0.0436	252	-0.0174	0.0385
Milk	178	0.0011	0.0154	184	0.1331	0.0637	170	0.0368	0.0651
Oats	480	0.0072	0.0308	252	0.3285	0.0692	252	-0.1315	0.0686
Orange juice	480	0.0076	0.0279	251	0.3218	0.1030	252	-0.0707	0.0699
Pork bellies	474	0.0074	0.0301	224	0.3710	0.0946	195	0.0074	0.0765
Rough rice	292	0.0020	0.0268	236	0.2744	0.1386	252	-0.0421	0.0786
Soybean meal	480	0.0095	0.0287	252	0.2401	0.0769	252	-0.0619	0.0533
Soybean oil	480	0.0088	0.0328	252	0.2295	0.0591	252	-0.0463	0.0638
Soybeans	480	0.0071	0.0240	252	0.2391	0.0779	252	-0.0438	0.0666
Sugar	480	0.0044	0.0445	252	0.3461	0.0987	252	-0.0527	0.0633
Wheat	480	0.0020	0.0259	252	0.2772	0.0800	252	-0.0248	0.0582
Non-agricultural commodities									
Blendstock gasoline	310	0.0181	0.0302	67	0.4065	0.1581	70	-0.0604	0.0190
Coal	113	0.0057	0.0318	27	0.3139	0.0686	19	-0.0108	0.0117
Copper	480	0.0107	0.0302	252	0.2788	0.1085	252	-0.0353	0.0709
Cotton	480	0.0053	0.0284	252	0.2609	0.0879	252	-0.0121	0.0698
Crude oil	333	0.0118	0.0330	252	0.3553	0.1222	252	-0.0072	0.0240
Gold	432	0.0039	0.0192	252	0.1663	0.0659	252	-0.0591	0.0938
Heating oil	386	0.0114	0.0297	252	0.3409	0.1049	252	-0.0337	0.0308
Lumber	480	-0.0023	0.0319	250	0.3333	0.0775	252	-0.0426	0.0916
Natural gas	249	-0.0025	0.0448	230	0.5341	0.1663	252	-0.0079	0.0369
Palladium	408	0.0115	0.0373	16	0.3638	0.0801	237	-0.1051	0.0903
Platinum	480	0.0071	0.0244	194	0.2157	0.0718	252	-0.1406	0.0726
Propane	252	0.0220	0.0446				114	-0.0421	0.0326
Silver	480	0.0078	0.0306	252	0.2580	0.1190	252	-0.1481	0.0473
Unleaded gas	250	0.0198	0.0297	192	0.3400	0.0980	192	-0.0310	0.0365

Figure 1. Basis Spread and Basis Factor Return over Business Cycles

The solid line is the basis factor return index (left axis); the dashed line is the basis spread (right axis); and the shaded area is the US recession period as determined by NBER. The index value is set to be 1 at the end of January 1972, and we take the logarithmic transformation of the index value.


log of index value (Jan 1972: 1)
basis spread