Regression Analysis on Asset Returns

By John Lee

Linear Regression Use Cases

- Understand relationship between a stock's historical prices and various factors like market indexes, interest rates, trading volumes, or company-specific financial indicators.

- Quantify the impact of risk factors to understand risks associated with investments.

- Attribute fundamental performances such as revenue, profits, or sales, to various factors like marketing expenditure, pricing strategies, market conditions, or macroeconomic indicators.

- Assess correlations between asset pairs to help with activities like hedging and pairs trading.

Multiple Linear Regression

The multiple linear regression model defines a linear functional relationship between one continuous outcome variable and MULTIPLE input variables that takes on the following form:

$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_n x_{tn} + \epsilon_t$, where

y_t represents the observed outcome at time t

β_0 represents the y-intercept, or y value when $x = 0$

β_{tn} represents the slope of the n-th explanatory variable at time t

ϵ_t represents the error term, which accounts for the total deviations from the linear hyperplane at time t

Multiple Linear Regression

- In reality, multiple regression is more frequently used than simple regression because a dependent variable is rarely explained by only one variable. For instance, it takes more than just the oil price to explain the inflation in the economy.
- Similar to the simple linear regression, we can use the OLS or MLE methods to estimate the hyperplane that best fit the data points.
- Since multiple regression is more complex than simple regression, it possesses lower bias and higher variance.

Predicting Russell 2000 Returns

We compiled a list of features for building our linear model for predicting the performance of a the Russell 2000 ETF (Ticker: IWM). These features include ETFs considering Environmental, Social and Governance (ESG) factors as well as economic indicators.

Technology Breakthrough	Social Change	Urbanization	Climate Change	Global Wealth
● IBLC ● IRBO ● IHAK	● IDNA ● IWFH ● BMED	● IFRA ● IGF ● EMIF	● ICLN ● IDRV ● IVEG	● CNYA

Predicting Russell 2000 Returns

- ETF data comes from *yfinance*; economics data comes from *pandas_datareader.*
- We gather these data dating between 2022-06-02 and 2023-06-31, a total of 248 days.

```
DatetimeIndex: 248 entries, 2022-06-02 to 2023-05-31
Data columns (total 17 columns):
 #    Column                     Non-Null Count    Dtype
---   ------                     --------------    -----
 0    BMED                       248 non-null      float64
 1    CNYA                       248 non-null      float64
 2    EMIF                       248 non-null      float64
 3    IBLC                       248 non-null      float64
 4    ICLN                       248 non-null      float64
 5    IDNA                       248 non-null      float64
 6    IDRV                       248 non-null      float64
 7    IFRA                       248 non-null      float64
 8    IGF                        248 non-null      float64
 9    IHAK                       248 non-null      float64
 10   IRBO                       248 non-null      float64
 11   IVEG                       248 non-null      float64
 12   IWFH                       248 non-null      float64
 13   IWM                        248 non-null      float64
 14   volatiliity_index          248 non-null      float64
 15   option_adjusted_spread     248 non-null      float64
 16   inflation_rate             248 non-null      float64
dtypes: float64(17)
```
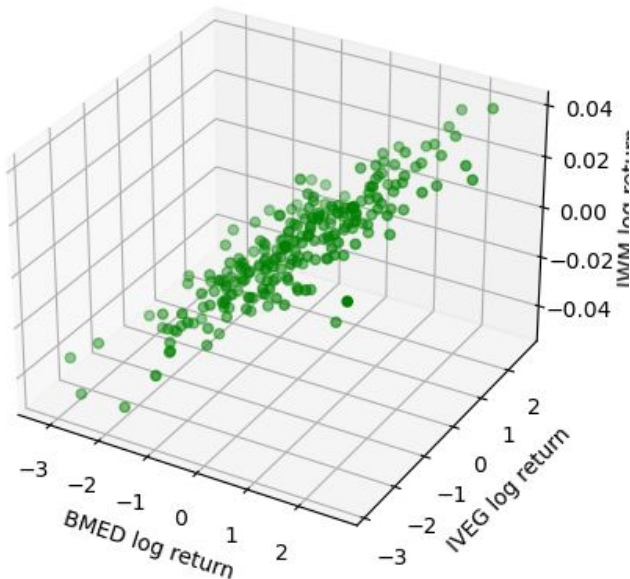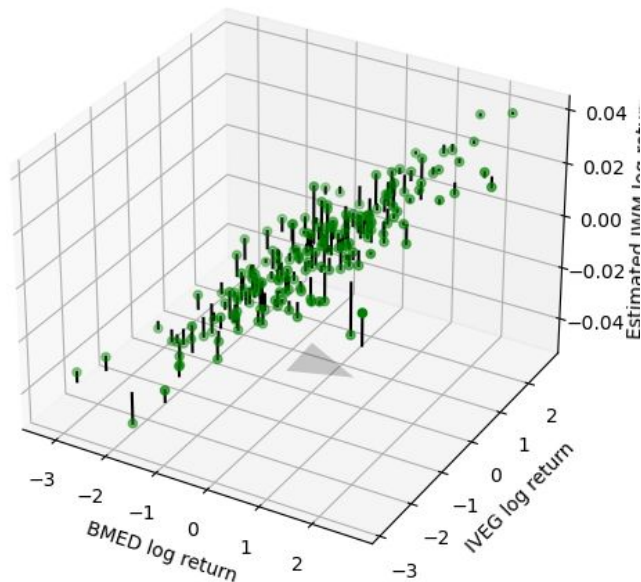
Multiple Linear Regression Fit

Our model has two independent variables, so its predictive formula looks something like this:

$$\text{IWM lo}\hat{\text{g}} \text{ return} = \hat{\beta_0} + \hat{\beta_1} \cdot (\text{BMED log return}) + \hat{\beta_2} \cdot (\text{IVEG log return})$$

BMED vs IVEG vs IWM Log Returns Before Fit

BMED vs IVEG vs IWM Log Returns After Fit

Fitting a Multiple Linear Regression

- This is a multiple linear regression with two independent variables
- Using OLS, we find that $\beta_0 = -0.0003$, $\beta_1 = 0.0077$, and $\beta_2 = 0.0081$ using OLS
- We fit the estimated coefficients into the regression model and arrived with :

$\hat{y}_t = -0.0003 + 0.0077 x_{t1} + 0.0081 x_{t2}$, where $\hat{y}_t$ denotes the predicted value at time t,.

where $\hat{y}_t$ denotes the predicted IWM log return at time t,

$\quad x_{t1}$ denotes BMED log return at time t,

$\quad x_{t2}$ denotes IVEG log return at time t.

BMED vs IVEG vs IWM Log Returns After Fit
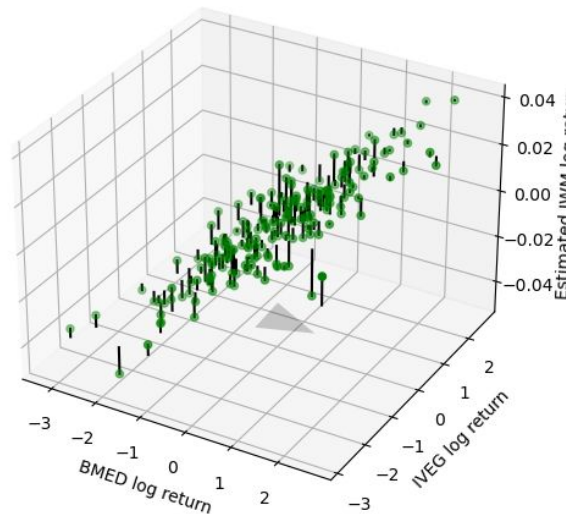
OLS Regression Result Summary

What can you say about this multiple linear regression model?

```
                         OLS Regression Results
==============================================================================
Dep. Variable:                    IWM   R-squared:                       0.874
Model:                            OLS   Adj. R-squared:                  0.873
Method:                 Least Squares   F-statistic:                     649.3
Date:                Wed, 07 Jun 2023   Prob (F-statistic):           7.01e-85
Time:                        12:17:05   Log-Likelihood:                 719.95
No. Observations:                 190   AIC:                            -1434.
Df Residuals:                     187   BIC:                            -1424.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0003      0.000     -0.723      0.471      -0.001       0.001
BMED           0.0077      0.001     13.699      0.000       0.007       0.009
IVEG           0.0081      0.001     14.378      0.000       0.007       0.009
==============================================================================
Omnibus:                        6.279   Durbin-Watson:                   1.952
Prob(Omnibus):                  0.043   Jarque-Bera (JB):                8.704
Skew:                           0.168   Prob(JB):                       0.0129
Kurtosis:                       3.993   Cond. No.                         2.36
==============================================================================
```

OLS Regression Result Summary

- **T-test:** p-value < 0.05, rejects null hypothesis and conclude that BMED and IVEG log returns are able to explain the variance of IWM log returns.

- **F-test:** p-value < 0.05, rejects the hypothesis and conclude that at least BMED or IVEG log returns can explain the variance of IWM log returns.

OLS Regression Result Summary

We fit the multiple linear regression model on the test data.

```
R-squared:  0.8496716652941945
Adjusted R-squared:  0.8429904059739364
MAE:  0.0045893760263598015
RMSE:  0.005849070813181978
```

- **R-squared:** the model explains 84.97% of the variation in IWM's log returns, without adjusting for penalty in feature numbers.

- **Adjusted R-squared:** the model explains 82.30% of the variation in IWM's log returns, adjusting for penalty in feature numbers.

- **MAE:** On average, the total absolute distance of the predicted log returns for IWM from their actual values is 0.0046.

- **RMSE:** The square root of the average of squared differences between the predicted IWM log returns and their actual values is 0.0058.

Regression Analysis on Asset Returns

By John Lee