INFO 7374 **Yizhen Zhao**
Summer I, 2018

Lecture 3 Machine Learning in Active Investment

A Short History of Investment Beliefs

- Pre-modern 1950 world, Fundamental Stock Selection - go for the "best" stocks and enjoy the results.
- 1960s Technical Analysis
- 1970s Efficient Market Hypothesis
- 1980s Market Anomalies: (1) Size effect
  (2) Calendar effect
  (3) Value irregularities
  (4) Momentum effect
  (5) Earnings Surprise Effect
  (6) Earnings Torpedo Effect

Quantitative Methods Involved

- Linear Regression
- Factor Selection/ Model Selection
- Model Combination

Active Alpha Strategies

- **Active Alpha Strategy** involves **two basic components:**
  1. The first is to build a multi-factor alpha model and then make return forecasts.
  2. The second is to choose optimal weights for individual stocks to maximize the information ratio (IR), which is the expected excess returns per unit of risk.

- **Alpha Factor Selection:**
  Alpha factors are mainly of three types:
  1. Value
  2. Momentum
  3. Quality.

  One could also include a long list of market anomalies in factors, using simultaneous equation, information-criterion or more advanced alpha modeling technique, such as contextual modeling.

- **Evaluation of Alpha Factors: Information Coefficient**

$$Information\ Coefficient\ (IC) = corr\ (f, r)$$

where f is a forecast of portfolio returns using alpha factors, and r represents actual portfolio returns. IC measures the cross-sectional correlation between the security return forecasts coming from a factor and the subsequent actual returns for securities.

1. Given other things equal, the higher the average IC for a factor is over time, the better the reward-to-risk ratio.
2. The more stable the IC over time, the better the result.
3. IC = Realized Portfolio Excess return/ dispersion of the forecasts (conviction) * dispersion of actual returns(opportunities)
4. An IC of 0.1 or higher on an annual basis is considered quite strong. If a factor f consistently has negative IC, we can just use -f as a factor.

- **Ultimate Objective** is to maximize

$$Information\ Ratio = \frac{mean(IC)}{std(IC)}$$

- **Optimal weights are chosen subject to different constraints:**

1. **A dollar-neutral strategy (a'i = 0)** makes the dollar amount of the long position equals the dollar amount of the short position. It doesn't take into account of the volatility (risk) of either side. Depending on volatility you either end up positively or negatively correlated with the market.
2. **A beta-neutral strategy (w'β = 0)** makes the weighted average beta of the shares in the long position equals the weighted average beta of shares in the short position so that the overall beta of the portfolio is zero. If CAPM is true, the beta-neutral fund should be totally insensitive to market movements. If you're long 130% in stocks with an average β of +0.3 and short 30% in stocks with an average β of +1.3, then your portfolio β is:

$$\beta = (130\% \times 0.3) + (-30\% \times 1.3) = 0.0.$$

3. **Sector-neutrality** means long and short positions are balanced by industry sectors.
4. **Factor-neutrality** means the exposure to factors such as the price of oil, the level of interest rates, or the rate of inflation is neutralized.
5. **A market -neutral strategy** eliminates the correlation to the market by hedging the long side with an equally risky (=same volatility) short side. E.g. A hedge of a stock portfolio with a short position on the S&P500 Future. The size of the short position is chosen in a way that the resulting strategy doesn't correlate with the S & P 500 anymore.

Regression Analysis

Ordinary Least Square Estimation

- In data analysis, the most commonly applied econometric tool is least-squares estimation, also known as **regression.** In a linear regression, the **dependent variable *Y*** is projected on a set of N predetermined **independent variables, *X*.**
- In the simplest bivariate case,

$$Y_t = \alpha + \beta X_t + \epsilon_t, \qquad t = 1, \dots, T$$

where α is called the **intercept**, or constant. β is called the **slope** and ϵ is called **error term**. Here, Y is the **regressand**, and X the **regressor** (as in **predictor**).

- In matrix form, Linear Regression

$$Y = X\beta + \epsilon$$

The best linear predictor of Y given x is $X\beta$, where β minimizes the mean squared error

$$S(\beta) = E(Y - X\beta)^2$$

The minimizer

$$\beta = \underbrace{argmin}_{\beta \in R^k} S(\beta)$$

The method is so-called **Ordinary Least Squares (OL**S), using these assumptions:

- The errors are independent of variables X: $cov(\epsilon, X) = 0$.
- The errors have constant variance: $var(\epsilon) = E(\epsilon'\epsilon) = \sigma^2$
- The errors are uncorrelated across observations t: $cov(\epsilon_{t-i}, \epsilon_{t-j}) = 0$.
- The errors are normally distributed: $\epsilon \sim N(0, \sigma^2)$.

- The OLS estimator is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

and errors

$$\hat{\epsilon} = Y - X\hat{\beta}$$

Once the parameter has been estimated, we can construct a forecast for y conditional on observations on x:

$$\hat{Y} = X\hat{\beta}$$

To interpret β, let's think about the bivariate case

$$cov\,(Y, X) = cov(\alpha + \beta X + \epsilon, X) = \beta cov(X, X) = \beta V(X)$$

because ϵ is uncorrelated with X. This shows that the population β is also

$$\beta(Y,X) = \frac{Cov(Y,X)}{V(X)} = \frac{\rho(Y,X)\sigma(Y)\sigma(X)}{\sigma^2(X)} = \rho(Y,X)\frac{\sigma(Y)}{\sigma(X)}$$

Bivariate Regression: Quality of Fit

- The regression fit can be assessed by examining the size of the residuals, obtained by

$$\widehat{\epsilon_t} = Y_t - \hat{Y}_t = Y_t - \hat{\alpha} - \hat{\beta}X_t$$

The estimated variance is

$$V(\hat{\epsilon}) = \frac{1}{(T-2)}\sum_{t=1}^{T} \widehat{\epsilon_t}^2$$

We divide by $T-2$ because the estimator uses two unknown quantities, $\hat{\alpha}$ and $\hat{\beta}$. Without this adjustment, it would be too small, or biased downward.

- Also, the estimated residuals must average to zero by construction

$$\frac{1}{T}\sum_{t=1}^{T} \hat{\epsilon} = 0$$

- The quality of fit can be assessed using a unit less measure called the **regression R-squared**, also called the **coefficient of determination.**

$$R^2 = 1 - \frac{SSE}{SSY} = 1 - \frac{\sum_{t=1}^{T} \widehat{\epsilon_t}^2}{\sum_{t=1}^{T}\left(Y_t - \overline{Y}\right)^2}$$

where SSE is the sum of squared errors (residuals, to be precise), and SSY is the sum of squared deviations of y around its mean. If the regression includes a constant, we always have $0 \leq R^2 \leq 1$.

- To interpret R^2, decompose the variance of Y_t: $Y_t = \alpha + \beta X_t + \epsilon$:

$$var(Y_t) = \beta^2 var(X_t) + var(\epsilon)$$

Divide by $var(Y_t)$:

$$1 = \frac{\beta^2 var(X_t)}{var(Y_t)} + \frac{var(\epsilon)}{var(Y_t)}$$

Because $R^2 = 1 - var(\epsilon_t)/var(Y_t)$, it is equal to $\frac{\beta^2 var(X_t)}{var(Y_t)}$, which is the contribution in the variation of y due to β and X.

$$R^2 = \frac{\beta^2 var(X_t)}{var(Y_t)} = \rho(Y,X)^2 \frac{var(Y)}{var(X)}\frac{var(X_t)}{var(Y_t)} = \rho(Y,X)^2.$$

Bivariate Regression: Hypothesis Testing

We can derive the distribution of the estimated coefficients, which is normal and centered around the true values. For the slope coefficients, $\hat{\beta} \sim N\left(\beta, var\left(\hat{\beta}\right)\right)$, with the variance given by

$$var\left(\hat{\beta}\right) = V(\hat{\epsilon}) \frac{1}{\sum_{t=1}^{T}(X_t - \bar{X})^2}$$

The associated test statistic

$$t = \frac{\hat{\beta}}{\sigma\left(\hat{\beta}\right)}$$

has a Student's t distribution.

Multivariate Regression

The above derivation is clearer in the multivariate case:

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \epsilon) = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon = \beta + (X'X)^{-1}X'\epsilon$$

$$var\left(\hat{\beta}\right) = \left(X'X\right)^{-1} X' var(\epsilon) X \left(X'X\right)^{-1} = var(\epsilon)\left(X'X\right)^{-1}$$

using

$$var(\epsilon) = \frac{1}{(T-N)} \sum_{t=1}^{T} \hat{\epsilon_t}^2$$

where the denominator is adjusted for the number of estimated coefficients N.

If we want to test whether the last m coefficients are jointly zero. Define $\widehat{\beta_m}$ as these grouped coefficients and $V_m(\hat{\beta})$ as their covariance matrix. We set up a statistic

$$F = \frac{\hat{\beta}'_m V_m\left(\hat{\beta}\right)^{-1} \hat{\beta}_m / m}{SSE/(T-N)}$$

which has an F-distribution with m and $T - N$ degrees of freedom. We would reject the hypothesis if the value of F is too large compared to critical values from tables.

The R-square mechanically increases as variables are added to the regression. with more variables, the variance of the residual term must be small, because this is in-sample fitting with more variables. Sometimes an adjusted R-squared is used

$$\bar{R}^2 = 1 - \frac{SSE/(T-N)}{SSY/(T-1)} = 1 - (1-R^2)\frac{T-1}{T-N}$$

This more properly penalizes for increasing the number of independent variables.

Note: When Matrices Are Singular

If the matrix $X'X$ is close to singular or badly scaled, the coefficient matrix (β) is most likely ill-conditioned. To address this problem, we usually use general/pseudo inverse.

we can define the **generalized inverse** or **g-inverse** as follows: Given an $m \times n$ matrix A, an $m \times n$matrix G is said to be a generalized inverse of A if

$$AGA = X$$

The Penrose conditions define different generalized inverses for $A \in R^{m\times n}$and $A^g \in R^{m\times n}$:

1. $AA^gA = A$
2. $A^gAA^g = A^g$
3. $(AA^g)^T = AA^g$
4. $(A^gA)^T = A^gA$

where A^T represents conjugate transpose. If A^g statisfies the first condition, then it is a **generalized inverse** of A. If it satisfies the first two conditions, then it is a **reflexive generalized inverse** of A. If it satisfies all four conditions, then it is the pseudo-inverse of A. A pseudo-inverse is sometimes called the **Moore–Penrose inverse**, after the works by E. H. Moore and Roger Penrose.

Pitfalls with Regression

Misspecification

Omitted variables: The true model has N variables but we use only a subset N_1. If the omitted variables are correlated with the included variables, the estimated coefficients will be biased.

Multicollinearity

Practical Consequences of Multicollinearity

- Although BLUE, $\hat{\beta}_{OLS}$ have large variances and covariances, making precise estimation difficult.
- Confidence intervals tend to be much wider, leading to acceptance of the "zero null hypothesis" more readily - insignificant.
- $t_{\beta_{OLS}}$ for one or more coefficients tend to be statistically insignificant
- Though insignificant, R^2 can be very high.
- $\hat{\beta}_{OLS}$ and their standard errors can be sensitive to small changes in the data.

Large Variances and Covariances of OLS Estimators

Due to Multicolinearity, the variances and covariances of $\hat{\beta}_2$ and $\hat{\beta}_3$ are given by

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{t=1}^{T}(X_{2t} - \bar{X}_2)^2} \cdot \frac{1}{1 - r_{23}^2}$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum_{t=1}^{T}(X_{3t} - \bar{X}_3)^2} \cdot \frac{1}{1 - r_{23}^2}$$

$$cov(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\sqrt{\sum_{t=1}^{T} x_{2t}^2 \sum_{t=1}^{T} x_{3t}^2}}$$

where r_{23} is the coefficient of correlation between X_2 and X_3.

The speed with which variances and covariances increase can be seen with the **Variance-Inflating Factor (VIF)**, which is defined as

$$VIF = \frac{1}{1 - r_{23}^2}$$

VIF shows how the variance of an estimator is inflated by the presence of collinearity. As r_{23}^2 approaches 1, the VIF approaches infinity. If there is no collinearity between X_2 and X_3, VIF will be 1. As the collinearity of regressors increases, VIF increases.

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{t=1}^{T} x_{2t}^2} VIF$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum_{t=1}^{T} x_{3t}^2} VIF$$

The results just discussed can be easily extended to the *k*-variable model. In such a model, the variance of the *j*-th coefficient, can be expressed as

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{t=1}^{T} x_{jt}^2} \frac{1}{1 - R_j^2}$$

where

- $\hat{\beta}_j$ = estimated partial regression coefficient of regressor X_j
- $R_j^2 = R^2$ in the regression of X_j on the remaining $(k - 2)$ regressions
  Note: except for intercept, there are $(k - 1)$ regressors in the k-variate regression model.
- $\sum_{t=1}^{T} x_{jt}^2 = \sum_{t=1}^{T}(X_{jt} - \overline{X}_j)_{jt}^2$

Detection of Multicollinearity

1. High R^2, but few significant t test statistics.
2. High pair-wise correlations among regressors.
3. Examination of partial correlations.
4. *Auxiliary Regressions*

   regress each X_i on the remaining X variables and compute the corresponding R^2, which we designate as R_i^2, each one of these regressions is called an auxiliary regression, auxiliary to the main regression of Y on $X's$. Then use the following F test to establish multicollinearity:

$$F_i = \frac{R^2_{x_i \cdot x_2 x_3 \cdots x_k}/(k-2)}{(1 - R^2_{x_i \cdot x_2 x_3 \cdots x_k})/(T-k+1)}$$

   follows the F distribution with $k-2$ and $T-k+1$ df. T stands for the sample size, k stands for the number of explanatory variables including the intercept term, and $R^2_{x_i \cdot x_2 x_3 \cdots x_k}$ is the coefficient of determination in the regression of variable X_i on the remaining X variables. If the computed F_i exceeds the critical F_i at chosen level of significance, it is taken to mean that the particular X_i is collinear with other $X's$. We have to decide whether we want to drop X_i from the model. If the computed F_i does not exceeds the critical F_i, we say that it is not collinear with other $X's$, in which case we may retain that variable in the model.

   Klien's Rule of Thumb: multicollinearity may be a troublesome problem only if the R^2 obtained from an auxiliary regression is greater than the overall R^2, which obtained from the regression of Y on all regressors.
5. Eigenvalues and condition index.

   Most software package uses eigen values and the conditional index to diagnose multicollinearity. From these eigenvalues, however, we can derive what is known as the condition number k defined as

$$k = \frac{Maximum\ eigenvalue}{Minimum\ eigenvalue}$$

Solution(Rule of Thumb)

- *A priori* information
- Combining cross-sectional and time-series data: pooling the data
- Dropping a variable(s) ans specification bias
- Transformation of Variables: first-difference, ratio transformation, logarithmic transformation.
- Additional or new data.
- Reduce collinearity in polynomial regressions.
- Other methods: factor analysis and principal components.

Heteroscedasticity

The Nature of Heteroscedasticity

Homoskedasticity

$$E(u_t^2) = \sigma^2 \qquad t = 1,2,\cdots,T$$

Heteroscedasticity

$$E(u_t^2) = \sigma_t^2 \qquad i = 1,2,\cdots,T$$

If σ_i^2 were known,

$$var(\hat{\beta}) = \frac{\sum_{t=1}^{T} x_{2t}^2 \, \sigma_t^2}{(\sum_{t=1}^{T} x_{2t}^2)^2}$$

whereas Homoskedasticity is

$$var(\hat{\beta}) = \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2}$$

- $\hat{\beta}_{OLS}$ is consistent, but inefficient,confidence interval too wide, $\hat{\beta}_{OLS}$ is insignificant.
- $\boldsymbol{\hat{\beta}_{GLS}}$ **is the BLUE estimator:** consistent and efficient,$var(\hat{\beta}_{GLS}) \leq var(\hat{\beta}_{OLS})$

Solution

- Idea: $Y_t = \beta_1 + \beta_2 X_t + u_t$, divide through by σ_t to obtain

  $$\frac{Y_t}{\sigma_t} = \beta_1 \left(\frac{1}{\sigma_t}\right) + \beta_2 \frac{X_t}{\sigma_t} + \frac{u_t}{\sigma_t}$$

  where

  $$var\left(\frac{u_t}{\sigma_t}\right) = E[\left(\frac{u_t}{\sigma_t}\right)^2] - \left[E\left(\frac{u_t}{\sigma_t}\right)\right]^2 = E[\left(\frac{u_t}{\sigma_t}\right)^2] = \frac{1}{\sigma_t^2} E(u_t^2) = \frac{\sigma_t^2}{\sigma_t^2} = 1$$

  GLS is OLS on the transformed variables that satisfy the standard least-square assumptions.

- Practice: **Generalized Least Squares (GLS) estimator** is infeasible since the matrix D is unknown. A feasible GLS (FGLS) estimator replaces the unknown D with an estimate

  $$\hat{D} = diag\{\hat{\sigma}_1^2, \cdots, \hat{\sigma}_T^2\}$$

We now discuss the estimation procedure

$$\sigma_t^2 = \alpha z_t$$

where z_{1t} is some $q \times 1$ function of X_t. Typically, z_{1t} are squares (and perhaps levels) of some (or all) elements of X_t. Often the functional form is kept simple for parsomony.

Let $\eta_t = u_t^2$, we could estimate α by OLS

$\tilde{\alpha} = (Z'Z)^{-1}Z'\eta \xrightarrow{p} \alpha$ and $\sqrt{n}(\tilde{\alpha} - \alpha) \xrightarrow{d} N(0, V_\alpha)$.

While u_t is not observed, we have the OLS residual $\hat{u}_t = Y_t - X_t\hat{\beta}$

As we can estimate $\sigma_t^2 = \alpha z_t$ by $\tilde{\sigma}_t{}^2 = \tilde{\alpha} z_t$
Suppose that $\tilde{\sigma}_i{}^2 > 0$ for all t. Then set

$$\tilde{D} = diag\,\{\tilde{\sigma}_1^2, \cdots \tilde{\sigma}_T^2\}$$

and

$$\tilde{\beta}_{GLS} = \left(X'\tilde{D}^{-1}X\right)^{-1} X'\tilde{D}^{-1}y$$

The is the feasible GLS, or FGLS estimator of β.
White type estimator

$$\tilde{V}_\beta = \left(\frac{1}{n}X\tilde{D}^{-1}X\right)^{-1}\left(\frac{1}{n}X\tilde{D}^{-1}\hat{D}\tilde{D}^{-1}X\right)\left(\frac{1}{n}X\tilde{D}^{-1}X\right)^{-1}$$

where $\hat{D} = diag\{\hat{u}_1^2, \cdots, \hat{u}_T^2\}$.

Consequences of Using OLS in the presence of Heteroscedasticity

- OLS estimation allowing for Heteroscedasticity: assume σ_t is known, still we have $var(\hat{\beta}_{GLS}) \leq var(\hat{\beta}_{OLS})$. Confidence intervals based on OLS will be unnecessarily larger. the t and F tests are likely to give us inaccurate results in that $var(\hat{\beta}_2)$ is overly large. A statistically insignificant coefficient may in fact be significant.
- OLS estimation disregarding Heteroscedasticity: we not only use $\hat{\beta}_{OLS}$ but also continue to use the usual (Homoskedasticity) variance formula =>***severe problem****:*

1. Running a standard OLS regression will yield $var(\hat{\beta}_{OLS}) = \frac{\sigma^2}{\sum_{t=1}^T x_t^2}$, which is a biased estimator of
2. $$var(\hat{\beta}) = \frac{\sum_{t=1}^T x_{2t}^2\,\sigma_t^2}{(\sum_{t=1}^T x_{2t}^2)^2}$$

3. It may overestimate or underestimate the latter, and in general we cannot tell whether the bias is positive (overestimation) or negative (underestimation). As a result, we can no longer rely on the conventionally computed confidence intervals and the conventionally employed t and F tests. Whatever conclusion we draw, may be very misleading.

Detection of Heteroscedasticity

- Nature of the Problem
- Graphical Method: plot $\hat{u}_t^2$ against $\hat{Y}_t$.
- **Park Test:** running the following regression

  $$\ln \hat{u}_t^2 = \alpha + \beta \ln X_t + v_t$$

  If β turns out to be statistically significant, it would suggest that Heteroscedasticity is present in the data. If it turns out to be insignificant, we may accept the assumption of Homoskedasticity.

- **Spearman's Rank Correlation Test**

  Spearman's rank correlation coefficient

$$r_s = 1 - 6[\frac{\sum {d_t}^2}{T(T-1)}]$$

  where d_t is the difference in the ranks assigned to two different characteristics of the *t*-th individual or phenomenon and n = number of individuals or phenomena ranked.

  Step 1. Fit the regression to data on Y and X and obtain the residuals $\hat{u}_t$.

  Step 2. Take the absolute value $|\hat{u}_t|$, tank both $|\hat{u}_t|$ and X_t (or $\widehat{Y}_t$) according to an ascending or descending order and compute Spearman's Rank Correlation Coefficient given previously.

  Step 3. Assume that the population rank correlation coefficient ρ_s is zero and $T > 8$, the significance of the sample r_s can be tested by t test as follows:

$$t - stat = \frac{r_s\sqrt{(T-2)}}{\sqrt{1-r_s^2}}$$

  with $df = n - 2$ (one-sided).

  if the computed t value exceeds the critical t value, we may accept the hypothesis of Heteroskedasticity; otherwise, we may reject it.

- **Goldfeld-Quandt Test**

  This method is applicable if one assumes that the heteroskedastic variance σ_i^2, is positively related to one of explanatory variables in the regression model.

  Consider the two-variable model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

  Suppose σ_i is positively related to X_i as

$$\sigma_i^2 = \sigma^2 X_i^2$$

  where σ^2 is a constant. It would mean σ_i^2 would be larger, the larger the value of X_i. To test this explicitly, take the following steps:

  Step 1. Order or rank the observations according to the values of X_i, beginning with the lowest X value.

  Step 2. Omit c central observations, where c is specified a priori, and divide the remaining $(\mathrm{n} - \mathrm{c})$observations into two groups, each of $(\mathrm{n} - \mathrm{c})/2$ observations.

  Step 3. Fit separate OLS regression to the first $(\mathrm{n} - \mathrm{c})/2$ observations and the last $(\mathrm{n} - \mathrm{c})/2$ observations, and obtain the respective residual sums of squares RSS_1 and RSS_2, RSS_1 representing the RSS from the regression corresponding to the smaller $\mathrm{X_i}$ values (the small variance group) and RSS_2 that from the larger $\mathrm{X_i}$ values (large variance group). These RSS each have

$$\frac{(\mathrm{n}-\mathrm{c})}{2} - \mathrm{k} \quad \text{or} \quad \frac{\mathrm{n}-\mathrm{c}-2\mathrm{k}}{2}\ \mathrm{df}$$

  where k is the number of parameters to be estimated, including intercept.

Step 4. Compute the ratio

$$\lambda = \frac{RSS_2/df}{RSS_1/df}$$

If u_i are assumed to be normally distributed (which we usually do), and if the assumption of Homoskedasticity is valid, then it can be shown that λ follows F distribution with numerator and denominator df each of $\frac{n-c-2k}{2}$.

If the computed λ is greater than the critical F at the chosen level of significance, we can reject the hypothesis of Homoskedasticity, that is we can say that Heteroskedasticity is very likely.

Limitation: The success of the Goldfeld-Quadt Test depends not only on the value of c (the number of central observations to be omitted), but also on identifying the correct X variable with which to order the observations. This limitation of this test can be avoided if we consider the **Breusch-Pagan-Godfrey Test.**

- **Breusch-Pagan-Godfrey Test**

Consider the *k*-variate linear regression model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

Assume that the error variance σ_i^2 is described as

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi}$$

some of all of the X′s can serve as Z′s. If $\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0, \sigma_i^2 = \alpha_1$, which is a constant. Therefore, to test whether σ_i^2 is homoskedastic, one can test the hypothesis that

$$\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$$

This is the basic idea behind Breusch-Pagan test**.** The test procedure is as follows.

Step 1. Estimate by OLS and obtain the residuals $\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_n$.

Step 2. Obtain $\tilde{\sigma}^2 = \sum \hat{u}_i^2/n$. This is the maximum likelihood (ML) estimator of σ^2 (The OLS estimator is $\sum \hat{u}_i/(n-k)$)

Step 3. Construct variables p_i defined as

$$p_i = \hat{u}_i^2/\tilde{\sigma}^2$$

Step 4. Regress p_i thus constructed on the $Z's$ as

$$p_i = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} + v_i$$

where v_i is the residual term of this regression.

Step 5. Obtain the ESS (explained sum of squares) and define

$$\Theta = \frac{1}{2} ESS$$

Assuming u_i are normally distributed, one can show that if there is Homoskedasticity and if the sample size n increases indefinitely (asymptotically), then

$$\Theta \sim_{asy} \chi^2_{m-1}$$

That is, Θ follows the chi-square distribution with $(m-1)$ degrees of freedom. If the computed Θ exceeds the critical χ^2 value at the chosen level of significance, one can reject the hypothesis of Homoskedasticity; otherwise one does not reject it.

- **White's General Heteroscedasticity Test**

Goldfeld-Quandt test requires reordering the observations with respect to X variable. BPG test is sensitive to the normality assumption. The general test of Heteroskedasticity proposed by White does not reply on the normality assumption and is easy to implement. The White Test proceeds as follows:

Step 1. Given the data, we estimate

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Step 2. We then run the following auxiliary regression

$$u_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i$$

The squared residuals from the original regression are regressed on the original X variables or regressors, their squared values and the cross products of the regressors. Obtain R^2 from this regression.

Step 3. Under the null hypothesis that there is no Heteroskedasticity, it can be shown that sample size (n) times the R^2 obtained from the auxiliary regression asymptotically follows the chi-square distribution with df equal to the number of regressors (excluding the constant term) in the auxiliary regression. That is

$$n \cdot R^2 \sim_{asy} \chi^2_{df}$$

In our example, $df = 5$ since there are five regressors in the auxiliary regression.

Step 4. If the chi-square value obtained above exceeds the critical chi-square value at the chosen level of significance, the conclusion is that there is Heteroskedasticity. If it does not exceed the critical chi-square value, there is no Heteroskedasticity. And

$$\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$$

Note: In cases where the White test statistic given above is statistically significant, Heteroskedasticity may not necessarily be the cause, but specification errors. In other words, the White test can be a test of pure Heteroskedasticity or specification error or both.

- no cross product, pure test of Heteroskedasticity
- use cross product, a test of both Heteroskedasticity and specification bias.

Autocorrelation

The term **autocorrelation** may be defined as "correlation between members of series of observations ordered in time. The classical linear regression model assumes that autocorrelation does not exist in the disturbances u_i:

$$E(u_i u_j) = 0$$

Autocorrelation:

$$E(u_i u_j) \neq 0$$

Why do correlation exist?

- Inertia
- Specification Bias: Excluded Variable Case.
- Specification Bias: Incorrect Functional Form.
- Cobweb Phenomenon
- Lags
- Manipulation of Data
- Data Transformation
- Nonstationarity

OLS Estimation in Presence of Autocorrelation

We assume disturbances, or error terms are generated by the following mechanism

$$u_t = \rho u_{t-1} + \varepsilon_t \quad -1 < \rho < 1$$

where ρ is the coefficient of autocovariance and ε_t is the stochastic disturbance term such that it satisfied the standard OLS assumption.

Return to our two variable regression model:

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

$$\hat{\beta}_{OLS} = \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} x_t^2}$$

and

$$var(\hat{\beta}_{OLS}) = \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2}$$

Now under AR(1) scheme,

$$var(\hat{\beta}_{AR}) = \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2}\left(\frac{1+r\rho}{1-r\rho}\right) = var(\hat{\beta}_{OLS})\left(\frac{1+r\rho}{1-r\rho}\right)$$

where r is the coefficient of autocorrelation.

The OLD variance estimator is biased. The magnitude depends on ρ.

BLUE Estimator in the Presence of Autocorrelation

The weighted least-square estimator (a special case of GLS) is efficient.

$$\hat{\beta}_2^{GLS} = \frac{\sum_{t=2}^{T}(x_t - \rho x_{t-1})(y_t - \rho y_{t-1})}{\sum_{t=1}^{T}(x_t - \rho x_{t-1})^2} + C$$

where C is a correction factor that may be disregarded in practice. Its variance is given by

$$var(\hat{\beta}_2^{GLS}) = \frac{\sigma^2}{\sum_{t=1}^{T}(x_t - \rho x_{t-1})^2} + D$$

where D too is a correction factor that may be disregarded in practice.

Consequences of Using OLS in Presence of Autocorrelation

- $\hat{\beta}_{OLS}$ is not BLUE.
- $var(\hat{\beta}_{AR})$ are likely to be wider than those based on the GLS procedure.
- To establish confidence intervals and to test hypotheses, one should use GLS and not OLS even though the estimators derived from the latter are unbiased and consistent.

OLS Estimation Disregarding Autocorrelation

- The residual variance $\hat{\sigma}^2 = \sum \hat{u}_t / (T-2)$ is likely to underestimate the true σ^2.
- As a result, we are likely to overestimate R^2.
- Even if $\hat{\sigma}^2$ is not underestimated, $var(\hat{\beta})$ may underestimate $var(\hat{\beta}_{AR})$, its variance under (first-order) autocorrelation, even though the latter is inefficient compared to $var(\hat{\beta}_2^{GLS})$.
- The usual *t* and *F* tests of significance are no longer valid, and if applied, are likely to give seriously misleading conclusions about the statistical significance of the estimated regression coefficients.

Detecting Autocorrelation

- Graphical Method
- The Runs Test (Geary Test)
- Durbin Watson d Test

  Durbin Watson dStatistic is defined as

  $$d = \frac{\sum_{t=2}^{T}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=2}^{T}\hat{u}_t^2}$$

  which is simply the ratio of the sum of squared differences in successive residuals to the RSS.

  It is important to note the assumptions underlying the *d* statistic:

  1. The regression model includes the intercept term.
  2. The explanatory variables *X* are nonstochastic, or fixed in repeated sampling.

3. The disturbances u_t are generated by the first order autoregressive scheme $u_t = \rho u_{t-1} + \varepsilon_t$
4. The error term u_t is assumed to be normally distributed.
5. The regression model does not include the lagged values of the dependent variable as one of the explanatory variables.
6. There are no missing observations in the data.

Now let us define

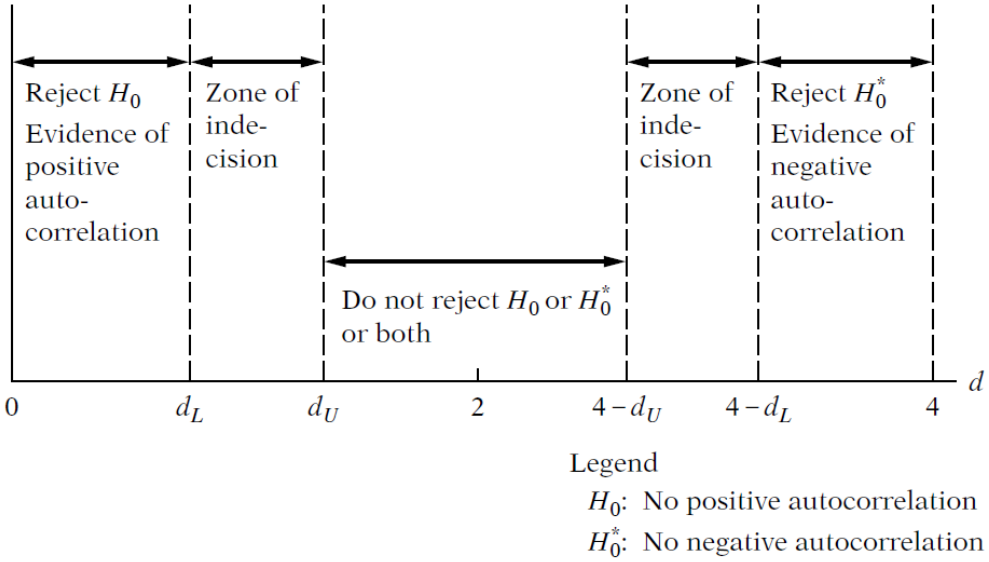
$$\hat{\rho} = \frac{\sum_{t=2}^{T}(\hat{u}_t \hat{u}_{t-1})}{\sum_{t=2}^{T} \hat{u}_t^2}$$

$$d \approx 2(1 - \hat{\rho})$$

Since $-1 \leq \rho \leq 1$ implies that

$$-1 \leq d \leq 1$$

If $\hat{\rho} = 0, \mathrm{d} = 2$, that is, if there is no serial correlation (of the first-order), d is expected to be about 2. Therefore, as a rule of thumb, if d is found to be 2 in an application, one may assume that there is no first-order autocorrelation, either positive or negative. If $\hat{\rho} = 1, \mathrm{d} = 0$, indicating perfect positive correlation in the residuals. Therefore, the closer d is to 0, the greater the evidence of positive serial correlation.



DURBIN–WATSON *d* TEST: DECISION RULES

Null hypothesis	Decision	If
No positive autocorrelation	Reject	$0 < d < d_L$
No positive autocorrelation	No decision	$d_L \leq d \leq d_U$
No negative correlation	Reject	$4 - d_L < d < 4$
No negative correlation	No decision	$4 - d_U \leq d \leq 4 - d_L$
No autocorrelation, positive or negative	Do not reject	$d_U < d < 4 - d_U$

- A General Test of Autocorrelation: The Breusch-Godfrey (BG) Test
  To avoid some of the pitfalls of the Durbin–Watson *d* test of autocorrelation, statisticians Breusch and Godfrey have developed a test of autocorrelation that is general in the sense that it allows for
  (1) non-stochastic regressors, such as the lagged values of the regressand;
  (2) higher-order autoregressive schemes, such as AR(1), AR(2), etc.; and
  (3) simple or higher-order **moving averages** of white noise error terms.

  BG Test, also known as the LM test, proceeds as follows:
  Let $Y_t = \beta_1 + \beta_2 X_t + u_t$. Assume that the error term u_t follows the pth-order autoregressive, *AR(p)* scheme as follows:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t$$

  where ε_t is a white noise error term.
  The null hypothesis H_0 to be tested is that

$$H_0: \rho_1 = \rho_2 = \cdots = \rho_p = 0$$

  The BG test involves the following steps:

  Step 1. Estimate $Y_t = \beta_1 + \beta_2 X_t + u_t$. by OLS and obtain the residuals $\hat{u}_t$.

  Step 2. Regress $\hat{u}_t$ on the original X_t and $\hat{u}_{t-1}, \hat{u}_{t-2}, \cdots, \hat{u}_{t-p}$, where the latter are the lagged values of residuals from Step 1:

$$\hat{u}_t = \alpha_1 + \alpha_2 X_t + \hat{\rho}_1 \hat{u}_{t-1} + \hat{\rho}_2 \hat{u}_{t-2} + \cdots + \rho \hat{u}_{t-p} + \varepsilon_t$$

  and obtain R^2 from this auxiliary regression.
  Step 3. If the sample size is large (technically, infinite), Breusch and Godfrey have shown that

$$(n - p)R^2 \sim \chi_p^2$$

  If $(n - p)R^2$ exceeds the critical chi-square value at the chosen level of significance, we reject the null hypothesis, in which case, at least one ρ is statistically significantly different from zero.

Model Selection

Information Criterion (AIC)

Akaike Information Criterion

$$\mathrm{AIC} = e^{2k/T}\frac{\sum \hat{u}_t^2}{T} = e^{2k/T}\frac{\mathrm{RSS}}{T}$$

for mathematical convenience

$$\ln \mathrm{AIC} = \left(\frac{2k}{T}\right) + \ln\left(\frac{\mathrm{RSS}}{T}\right)$$

Bayesian Information Criterion

$$\mathrm{BIC} = T^{k/T}\frac{\sum \hat{u}_t^2}{T} = n^{k/T}\frac{\mathrm{RSS}}{T}$$

for mathematical convenience

$$\ln \mathrm{BIC} = \left(\frac{k}{T}\right)\ln T + \ln\left(\frac{\mathrm{RSS}}{T}\right)$$

Shrinkage Methods

By retaining a subset of the predictors and discarding the rest, subset selection produces a model that is interpretable and has possibly lower prediction error than the full model. However, because it is a discrete process, variables are either retained or discarded—it often exhibits high variance,and so doesn't reduce the prediction error of the full model. Shrinkagemethods are more continuous, and don't suffer as much from highvariability.

Ridge Regression

Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin}\{\sum_{t=1}^{T}\left(Y_t - \beta_0 - \sum_{j=1}^{k} x_{tj}\beta_j\right)^2 + \lambda\sum_{j=1}^{k}\beta_j^2\}$$

Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the large the value of λ, the greater the amount of shrinkage. The coefficients are shrunk toward zero (and each other). The idea of penalizing by the sum-of-squares of the parameters is also used in neural networks, where it is known as weight decay.

The ridge solutions are not equivariant under scaling of the inputs, andso one normally standardizes the inputs before solving. Writing the criterion in matrix form,

$$RSS(\lambda) = (Y - X\beta)'(Y - X\beta) + \lambda\beta'\beta$$

The Ridge regression solution are easily seen to be

$$\hat{\beta}^{ridge} = (X'X + \lambda I)^{-1} X'Y$$

where I is the $k \times k$ identity matrix. With the choice of quadratic penalty $\beta'\beta$, the ridge regression solution is again a linear function of Y. The solution adds a positive constant to the diagonal of $X'X$ before inversion. This makes the problem nonsingular, even if $X'X$ is not full rank and was the main motivation for Ridge regression when it was first introduced. In univariate case, $\hat{\beta}^{ridge} = \hat{\beta}/(1+\lambda)$.

Choice of λ: suppose $Y_t \sim N(\beta_0 + X_t'\beta, \sigma^2)$ and the parameters β_j are each distributed as $N(0, \tau^2)$, independently of one another; then $\lambda = \sigma^2/\tau^2$.

LASSO Regression

The lasso is a shrinkage method like ridge, with subtle but important differences. The lasso estimate is defined by

$$\hat{\beta}^{LASSO} = \underset{\beta}{argmin}\left\{\frac{1}{2}\sum_{t=1}^{T}\left(Y_t - \beta_0 - \sum_{j=1}^{k} X_{tj}\beta_j\right)^2 + \lambda\sum_{j=1}^{k}|\boldsymbol{\beta_j}|\right\}$$

Here, the L_2 Ridge penalty $\sum_{j=1}^{k}\beta_j^2$ is replaced by L_1 Lasso penalty $\sum_{j=1}^{k}|\boldsymbol{\beta_j}|$. This latter constraint makes the solutions nonlinear in the Y_t, and there is no closed form expression as in ridge regression. Computing the lasso solution is a quadratic programming problem.

L_1 regularization has taken on a life of its own, leading to the development of the field ***compressed sensing*** in the signal-processing literature.

Least Angle Regression (LARS)

LAR is intimately connected with the lasso, and in fact provides an extremely efficient algorithm for computing the entire lasso path. It can be viewed as a kind of "democratic" version of forward stepwise regression.

Forward stepwise regression builds a model sequentially, adding one variable at a time. At each step, it identifies the best variable to include in the active set, and then updates the least squares fit to include all the active variables.

Least angle regression uses a similar strategy, but only enters "as much" of a predictor as it deserves. At the first step it identifies the variable most correlated with the response. Rather than fit this variable completely, LAR moves the coefficient of this variable continuously toward its least-squares value (causing its correlation with the evolving residual to decrease in absolute

value). As soon as another variable "catches up" in terms of correlation with the residual, the process is paused. The second variable then joins the active set, and their coefficients are moved together in a way that keeps their correlations tied and decreasing. This process is continued until all the variables are in the model, and ends at the full least-squares fit.

Algorithm: Least Angle Regression

Step 1. Standardize the predictors to have mean zero and unit norm. Start with the residual

$$\mathrm{r} = \mathrm{Y} - \overline{\mathrm{Y}}, \beta_1, \beta_2, \cdots, \beta_\mathrm{k} = 0$$

Step 2. Find the predictor X_j most correlated with r.

Step 3. Move β_j from 0 towards its least squares coefficient $\langle X_j, r \rangle$ until some other competitor X_k has as much correlation with the current residual as does X_j.

Step 4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $\langle X_j, X_k \rangle$ until some other competitor X_l has as much correlation with the current residual.

Step 5. Continue in this way until all k predictors have been entered. After $\min(T-1, p)$ steps, we arrive at the full least-squares solution.

Suppose A_k is the active set of variables at the beginning of the k-thstep, and let β_{A_k} be the coefficient vector for these variables at this step; there will be $k-1$ nonzero values, and the one just entered will be zero. If $\mathrm{r_k} = \mathrm{Y} - \mathrm{X_{A_k}}\beta_{A_k}$ is the current residual, then the direction for this step is

$$\delta_k = (X_{A_k}{}' X_{A_k})^{-1} X_{A_k} r_k$$

The coefficient profile then evolves as $\beta_{A_k}(\alpha) = \beta_{A_k} + \alpha \cdot \delta_k$. The direction chosen in this fashion is to keep the correlation tied and decreasing. If the fit vector at the begging for this step is $\hat{f}_k$, then it evolves as $\hat{f}_k(\alpha) = \hat{f}_k + \alpha \cdot u_k$, where $u_k = X_{A_k}\delta_k$ is the new direction. The name least angle arises from a geometrical interpretation of this process; u_k makes the smallest (and equal) angle with each of the predictors in A_k

The LAR algorithm is extremely efficient, requiring the same order of computation as that of a single least squares fit using the p predictors. Least angle regression always takes p steps to get to the full least squares estimates. The lasso path can have more than p steps, although the two are often quite similar.

Elastic Net

Elastic Net Penalty is a different compromise between ridge and lasso. The elastic-net selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge. It also has considerable computational advantages over the L_q penalties.

$$\lambda \sum_{j=1}^{k} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|)$$

Principle Components Regression

To obtain principal component

Step 1. Calculate the variance-covariance matrix of X, S.

Step 2. Make the eigen value decomposition

$$|S - \lambda I| = 0$$

λ is the eigen value. Eigen values are typically ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$

Step 3. Eigen vector can be obtained by solving a_j from $(S - \lambda_j I)a_j = 0, \quad j = 1, \cdots, d,$ subject to the condition that the set of eigen vectors are orthonormal. Mathematically, this can be expressed as

$$a_i^T a_i = I$$

$$a_i^T a_j = 0$$

Step 4. The j-th Principal Component is

$$z_j = a_j^T (X - \bar{X})$$

while the principal component scores are contained in the matrix $\boldsymbol{Y = X\Lambda}$.

Supervised Principal Components

We want to encourage principal component analysis to find linear combinations of features that have high correlation with the outcome. To do this, we restrict attention to features which by themselves have a sizable correlation with the outcome. We use Supervised principal components. Supervised principal components is useful for linear regression. Its most interesting applications may be in survival studies.

Algorithm: Supervised Principal Components.

1. Compute the standardized univariate regression coefficients for theoutcome as a function of each feature separately.
2. For each value of the threshold θ from the list$0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_k$:

(a) Form a reduced data matrix consisting of only those featureswhose univariate coefficient exceeds θ in absolute value, andcompute the first m principal components of this matrix.

(b) Use these principal components in a regression model to predict the outcome.

3. Pick θ (and m) by cross-validation.

Examples of Machine Learning in Active Investment

Example 1: **Cross Hedging Using Regression**

- A short hedge is a hedge that involves s short position in futures contracts. A short hedge is appropriate when the hedge already owns an asset and expects to sell it at some time in the future (expects price to fall). A short hedge can also be used when an asset is not owned right now but can be owned in the future.
- Hedges that involves taking a long position in a futures contract are known as long hedges. A long hedge is appropriate when a company knows it will have to purchase a certain asset in the future and wants to lock in a price now (expects price to rise).
- Cross hedging occurs when two assets are different. The hedge ratio is the ratio of the size of the position taken in futures contract to the size of the exposure. When the asset underlying the futures contract is the same as the asset being hedged, it is natural to use a hedge ratio of 1.0.
- When cross hedging is used, the hedger should choose a value for the hedge ratio that minimizes the variance of the value of the hedged position.
- Define:

  ΔS: Change in the spot price, S, during a period of time equal to the life of the hedge.

  ΔF: Change in the futures price, F, during a period of time equal to the life of the hedge.
- **Minimum Variance Hedge Ratio:** h^*

$$\Delta S_t = h^* \cdot \Delta F_t + \varepsilon_t \text{ and}$$

$$h^* = (F'F)^{-1}F'S = \rho\frac{\sigma_S}{\sigma_F}$$

  h^* is the average change in S for a particular change in F. $\sigma_s$is the standard deviation of ΔS, σ_F is the standard deviation of ΔF, and ρ is the correlation between the two.
- **The Hedge Effectiveness** can be defined as the proportion of the variance that is eliminated by hedging. This is the R^2 from the regression of ΔS against ΔF and equals ρ^2.
- **Optimal Number of Contracts:** To calculate the number of contracts that should be used in hedging, define:

  Q_S: size of position being hedged.

  Q_F: size of one futures contract (units).

  N^*: optimal number of futures contracts for hedging

$$N^* = h^*\frac{Q_s}{Q_F}$$

  Note: For \$1 move in futures contract, spot price moves \$$h^*$.

  Example: An airline expects to purchase 2 million gallons of jet fuel in 1 month and decide to use heating oil futures for hedging. Each heating oil contract traded on NYMEX is on 42000 gallons of heating oil. Given $\sigma_F = 0.0313$, $\sigma_S = 0.0263$, and $h = 0.928$.

$$h^* = 0.928 \times \frac{0.0263}{0.0313} = 0.7777$$

The optimal number of heating oil contracts is

$$0.7777 \times \frac{2000000}{42000} = 37.03$$

Example 2: **Hedging Non-Parallel Yield Curve Shift**

Principal Component Analysis Revisit

Three Properties

1. $\sum_{i=1}^{N} \sigma_{PC_i}^2 = \sum_{i=1}^{N} \sigma_i^2$ (sum of variances of rates)
2. $\text{corr}(PC_i, PC_j) = 0$: while individual rates can be highly correlated with one another.
3. The 1st PC explains the largest fraction of the sum of the variances of rates. The 2nd PC explains the next largest fraction.

PCA was mainly used in empirically based hedges for large portfolios. Hedging for single position mainly relies on regression based hedge.

	Factor Loadings for US Treasury Data									
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
3m	0.21	-0.57	0.50	0.47	-0.39	-0.02	0.01	0.00	0.01	0.00
6m	0.26	-0.49	0.23	-0.37	0.70	0.01	-0.04	-0.02	-0.01	0.00
12m	0.32	-0.32	-0.37	-0.58	-0.52	-0.23	-0.04	-0.05	0.00	0.01
2y	0.35	-0.10	-0.38	0.17	0.04	0.59	0.56	0.12	-0.12	-0.05
3y	0.36	0.02	-0.30	0.27	0.07	0.24	-0.79	0.00	-0.09	-0.00
4y	0.36	0.14	-0.12	0.25	0.16	-0.63	0.15	0.55	-0.14	-0.08
5y	0.36	0.17	-0.04	0.14	0.08	-0.10	0.09	-0.26	0.71	0.48
7y	0.34	0.27	0.15	0.01	0.00	-0.12	0.13	-0.54	0.00	-0.68
10y	0.31	0.30	0.28	-0.10	-0.06	0.01	0.03	-0.23	-0.63	0.52
30y	0.25	0.33	0.46	-0.34	-0.18	0.33	-0.09	0.52	0.26	-0.13

- PC1: corresponds to a parallel shift in the yield curve.
- PC2: corresponds to rotation or change of slope of the yield curve. Rates between 3 months and 2 years move in one direction; rates between 3 years and 30 years move in the other direction.
- PC3: corresponds to a "bowing" of yield curve. Rates at the short end and long end of the yield curve move in one direction; rates in the middle move in the other direction.

The interest rate move for a particular factor is known as **factor loading**. The quantity of a particular factor in the interest rate changes on a particular day is known as **factor score** of that day.

Standard Deviation of Factor Scores

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
	17.49	6.05	3.10	2.17	1.97	1.69	1.27	1.24	0.80	0.79

As variance of each rate series after standardization is 1, if PC1 moves by 1 bps,

- 1yr rate moves by $0.21 \times 17.49 = 3.6729$ bps
- 2yr rate moves by $0.35 \times 17.49 = 5.8$ bps

Findings

- The importance of a factor is measured by the standard deviation of its factor score.
- Factors are chosen so that factor scores are uncorrelated.
- The variances of the factor scores have the property that they add up to the total variance of the data.

  The total variance of the original data

  $$17.49^2 + 6.05^2 + 3.10^2 + 2.17^2 + 1.97^2 + \cdots + 0.79^2 = 367.9$$

  It can be seen that the first factor accounts for

  $$\frac{17.49^2}{367.9} = 83.1\%$$

  of the variance in the original data.

  The first two factors account for

  $$\frac{17.49^2}{367.9} + \frac{6.05^2}{367.9} = 93.1\%$$

  of the variance in the data.

  The third factor accounts for further

  $$\frac{3.10^2}{367.9} = 2.8\%$$

  of the data.
- Most of the interest rates move is accounted for by the first two to three factors.

Example: Using PCA to calculate VaR

	Change in Portfolio Value for a 1 Basis Point Rate Move			
1year rate	2 year rate	3 year rate	4 year rate	5 year rate
+10	+4	-8	-7	+2

The portfolio exposure to the first factor is

$$10 \times 0.32 + 4 \times 0.35 - 8 \times 0.36 - 7 \times 0.36 + 2 \times 0.36 = -0.08$$

and exposure to the second factor is

$$10 \times (-0.32) + 4 \times (-0.10) - 8 \times 0.02 - 7 \times 0.14 + 2 \times 0.17 = -4.40$$

The changes in the portfolio value is, to a good approximation, given by

$$\Delta P = -0.08 \cdot f_1 - 4.40 \cdot f_2$$

The factor scores are uncorrelated and have the standard deviations given above. The standard deviation of ΔP is therefore

$$\sqrt{0.08^2 \times 17.49^2 + 4.40^2 \times 6.05^2} = 26.66$$

The 1-day 99% VaR is $26.66 \times 2.33 = 62.12$.

Note: A principal components analysis can be used to identify factors describing movements in the indices and the most important of these can be used to replace the market indices in a VaR analysis.

Example 3: **Generate Trading Signals Using Price Momentum**

TA Indicators for Simple Price Momentum Signals

Moving Averages

Double Crossover

- A buy signal is produced when the shorter average crosses above the longer.
- A sell signal is produced when the shorter averages moves below the longer average.
- Pairs Choice: 5day - 20day, 10day - 50day

Triple Crossover

- 4-9-18 method is used mainly in futures trading. 5-10-20 day moving averages are widely used in commodity circles.
- A buying alert takes place in a downtrend when the 4 day crosses above both the 9 and the 18.
- A confirmed buying signal occurs when the 9 day then crosses above the 18.
- When the uptrend reverses to downside, the first thing that should take place is that the shortest (and the most sensitive) average – the 4 day - dips below the 9 day and the 18 day. This is only a selling alert. Some traders, however, might use that initial crossing as reason enough to begin liquidating long positions. Then, if the next longer average - the 9 day – drops below the 18 day, a confirmed sell short signal is given.

Moving Average Envelope

- Percentage envelopes can be used to help determine when a market has gotten overextended in either direction. They tell us when prices have strayed too far from their moving averages line.
- Short term traders often use 3% envelopes around a simple 21 day moving average. When prices reach one of the envelopes (3% from the average), the short-term trend is considered to be overextended. For long range analysis, some possible combinations include 5% envelopes around a 10-week average or a 10% envelope around a 40-week average.

Bollinger Bands

- Two trading bands are placed around a moving average similar to the envelope technique.
- Bolling Bands are placed two standard deviations above and below the moving average, which is usually 20 days (*What's the underlying assumption here? Normality*).
- Using two standard deviations ensures that 95% of the price data will fall between the two trading bands. Each touch of the lower band signaled an important market bottom and a buying opportunity.

Oscillators and Contrary Opinion

Momentum

- Market momentum is measured by continually taking price differences for a fixed time interval. The formula for momentum is

$$M = V - V^X$$

where V is the latest closing price and V^X is the closing price X days ago.

- While the 10 day momentum is a commonly used time period for reasons discussed later, any time period can be employed. A shorter time period produces a more sensitive line with more pronounced oscillations. A longer number of days (such as 40 days) results in a much another line in which the oscillator swings are less volatile.
- If prices are rising and the momentum line is above zero line and rising, this means the upward trend is accelerating. If the up-slanting momentum line begins to flatten out, this means the new gains being achieved by the latest closes are the same as the gains 10 days earlier. While prices may still be advancing, the rate of ascent (or the velocity) has leveled off. When the momentum line begins to drop toward the zero line, the uptrend in prices is still in force, but at a decelerating rate. The uptrend is losing momentum.
- The momentum chart has zero line. One could use the crossing of the zero line to generate buy and sell signals. A crossing above the zero line would be a buy signal, and a crossing below the zero line, is a sell signal.

The Relative Strength Index (RSI)

$$RSI = 100 - \frac{100}{1 + RS}$$

$$RS = \frac{Average\ of\ x\ days'\ up\ closes}{Average\ of\ x\ days'\ down\ closes}$$

- The 80 level usually becomes the overbought level in bull markets and 20 level the oversold level in bear markets.
- 14 days are usually used in the calculation.
- The 50 level is the RSI midpoint value, and will often act as support during pullbacks and resistance during bounces. Some traders treat RSI crossings above and below the 50 level as buying and selling signals respectively.

Moving Average Convergence/Divergence (MACD)

- The faster line (called the MACD line) is the difference between two exponentially smoothed moving averages of closing prices (usually the last 12 and 26 days or weeks). The slower line (the signal line) is usually a 9 period exponentially smoothed average of the MACD line.
- Most traders utilize the default values of 12, 26, and 9 in all instances. That would include daily and weekly values.
- The actual buy and sell signals are given when the two line cross. A crossing by the faster MACD line below the slower is a sell signal. In that sense, MACD resembles a dual moving average crossover method.
- MACD values also fluctuate above and below a zero line. That's where it begins to resemble an oscillator. An overbought condition is present when the lines are too far above the zero line. An oversold condition is present when the lines are too far below the zero line. The best buy signals are given when prices are well below the zero line (oversold).
- The two MACD lines can be turned into an MACD histogram. The histogram has a zero line of its own. When the MACD lines are in positive alignment (faster line over the slower), the histogram is above its zero line. Crossings by the histogram above and below its zero line coincide with actual MACD crossover buy and sell signals.
- The real value of the histogram is spotting when the spread between the two lines is widening or narrowing. When the histogram is over its zero line (positive) but starts to fall toward the zero line, the uptrend is weakening. When the histogram is below its zero line (negative) and starts to move upward toward the zero line the downtrend is losing its momentum.
- Histograms are best used for spotting early exit signals from existing positions. It is much more dangerous to use histogram turns as an excuse to initiate new positions against the prevailing trend.

On Balance Volume (OBV)

On Balance Volume (OBV) is a momentum indicator that uses volume flow to predict changes in stock price. The underlying assumption is when volume increases sharply without a significant change in the stock price, the price will eventually jump upward, and vice versa.

$$OBV_t = OBV_{t-1} + Volume_t \cdot I(Close_t > Close_{t-1}) - Volume_t \cdot (Close_t \leq Close_{t-1})$$

Note: All above signals are just reference values. You may alter them to fit your own trading needs and assets.

Signal Modeling

- Given a threshold θ, the signal is

$$signal_t = \begin{cases} 1 & buy & for\ P_t < -\theta \\ 0 & hold & for\ -\theta < P_t < \theta \\ -1 & sell & for\ P_t > \theta \end{cases}$$

Neural Networks

Assume $s_t^{n_1,n_2}$ is the trading signals generated from the short n_1 and the long n_2 moving averages. Under general regularity conditions, a sufficiently complex single hidden layer feed-forward network can approximate any number of a class of functions to any desired degree of accuracy.

The Linear Test Regression

$$r_t = \alpha + \sum_{i=1}^{p} \beta_i r_{t-i} + \sum_{i=1}^{p} \eta_i s_{t-i}^{n_1,n_2} + \epsilon_t \; \epsilon_t \sim ID(0, \sigma_t^2)$$

Single Layer Feed-forward Network Model

$$r_t = \alpha + \sum_{i=1}^{p} \beta_i r_{t-i} + \sum_{j=1}^{d} \eta_j G(\alpha_j + \sum_{i=1}^{p} \gamma_i s_{t-i}^{n_1,n_2}) + \epsilon_t \; \epsilon_t \sim ID(0, \sigma_t^2)$$

where G is the **activation function** which is chosen to be a sigmoidal function:

$$G(x) = \frac{1}{1 + e^{-\alpha x}}$$

Single Layer Feed-forward Network Model with lagged returns alone:

$$r_t = \alpha + \sum_{i=1}^{p} \beta_i r_{t-i} + \sum_{j=1}^{d} \beta_j G(\alpha_j + \sum_{i=1}^{p} \gamma_i r_{t-i}) + \epsilon_t \; \epsilon_t \sim ID(0, \sigma_t^2)$$

d is the total types of signals you want to enclose in the prediction.
p is the total numbers of lags you choose to enclose in the information set of prediction.

Random Forecast

Random forest is based on decision trees, where each node considers a certain feature chosen according to its improvement on a given loss function. Since decision trees tend to over-fit their training set, random trees add a random part in that several decision trees are generated from bootstrap samples of data and then averaged. In the process, the variance is greatly reduced at the price of a slight increase in bias, producing a more robust model.

We used the `RandomForestClassifier`for the Python `scikit learn library.` Some of the parameters of importance are

_ `num_estimators:` the number of trees to be generated. Since the decrease in variance varies slower than the increase in computational power, an average number of 100 is enough.

_ `criterion:` the loss function to estimate to quality of the splits when selecting features. Both the Gini criterion and the entropy have been tested with no real change.

_ `max_features:` the maximum number of features to consider. We chose it be equal to `n_features`to explore a maximum of possibilities.

_ `max_depth:` the maximum depth of the tree, chosen to be default so that the tree is expanded until there aren't enough samples to fill the leaves.

_ `min_samples_split`: the minimum amount of samples required to split an internal node. We chose to be twice the following parameter `min_samples_leaf` for the leaves to bepossible.

_ `min_samples_leaf:` the minimum amount of samples required to create a leaf of the tree.The default value is 1, however a value of 3 is preferred. Indeed, a smaller value will makethe classifier more liberal, which already is the case. Choosing a slightly higher value for thisparameter allows to counteract this fact and avoid excessive overfit.

The random forest model is more liberal and gives better prediction when there actually is a trend in the evolution of the price, while having poorer results when the market is only fluctuating.

Stochastic Gradient Descent

The third machine learning algorithm we decided to test on our data was Linear Stochastic Gradient Descent (SGD). For example, to train our linear model on our data, one can iteratively fit one hundred linear models on 4830 data points in 15 day rolling window. The first 4000 of these points were part of our training set and the other 830 points were part of our validation set.

SGD uses gradient descent to find the minimum or maximum of a given function. In our case, we sought to minimize the log loss function on our data. The log loss function is a classification loss function used as an evaluation metric. Specifically, we may try to classify the signals as +1 (price going up), 0 (price staying neutral), or -1 (price going down), the log loss function quantifies the accuracy of our classifier by penalizing the false classifications our linear model

makes. Using SGD allows us to select the linear model that minimizes the number of incorrect predictions we make via the log loss function. The algorithm works as follows:

1. Choose an initial vector of parameters ω and learning rate η
2. Repeat until shuffle examples in the training set.
3. For $i = 1,2, \cdots, n$ do:

$$\omega := \omega - \eta \Delta Q_i(\omega)$$

In your code, you may imported the linear SGD algorithm from the scikitlearn library in python. Please note the following:

- $\omega$is a vector of zeros of size n
- n is the length of our training set (in above example $\omega = 4000$)
- The stochastic gradient descent process is repeated 500 times to find the minimum of the log loss objective function
- $Q_i(\omega)$ is the log loss objective function (the gradient of this function is taken in the formula)
- $\eta$the learning rate is internally optimized by the scikit learn library