① Fama-Bacbeth　　　Option In the money : 赚钱

Run T cross-sectional regressions

$$R^e_{i,t} = \lambda_{0,t} + \lambda_{1,t} \ln BM_{i,t-1} + e_{i,t}$$

Save all the T estimated $\lambda_{1,t}$ and obtain

$$\hat{\lambda}_1 = \frac{1}{T} \sum_{t=1}^{T} \lambda_{1,t}$$

$$Var(\hat{\lambda}_1) = var_T(\lambda_{1,t}) / T$$

$$\lambda_{1,t} = \frac{1}{N_t} \frac{\underbrace{\ln BM'_{t-1}}_{(1\times N)} - E_i(\ln BM_{i,t-1})}{\underbrace{var_i(\ln BM_{i,t-1})}_{\text{across stock at a time}}} \underbrace{R^e_t}_{(N\times 1)}$$

Weight of Fama-McB　　　$(\sum w_{i,t} = 0)$

$$w_{i,t-1} = \frac{1}{N} \frac{(X_{i,t-1} - E_i[X_{i,t-1}])}{Var_i[X_{i,t-1}]}$$

$\lambda_{1,t} = \sum_{i=1}^{N} w_{i,t-1} R^e_{i,t}$ is (return of) a long-short portfolio

t-stat　$t = \dfrac{\hat{\lambda}_1}{\sqrt{Var_T(\lambda_{1,t})}} \sqrt{T}$　　$\hat{\lambda}_1 \to$ ave. excess return; denominator: standard deviation

Fama-MacBeth Regression test is a test of whether a traded long-short portfolio has a high Sharpe Ratio.

② Panel Regression　　　Panel: Two dimensions

typically cross-section and time series

Balanced Panel: N observations in cross-section for each t

Unbalanced Panel: For each t, only a subset $(N(t) < N)$

$$Cov(X,Y) = E[(X-E[X])(Y-E[Y])] = E[XY] - E[X]E[Y]$$

$$= \frac{1}{n} \sum_{i=1}^{n} (X_i - E[X])(Y_i - E[Y])$$

Implicit Assumption :

Slope coefficients do not vary over time or across firms

Intercepts can be varied : over time / across firms

Canonical Panel regression. $\longrightarrow$ firm-fixed effects

$$y_{i,t} = \delta_t + \theta_i + \beta' X_{i,t} + \varepsilon_{i,t}$$

time-fixed effects — effect is the same

Clustering for Panel Regression

1. Firm's shocks correlated within each year but not across years
2. Cross-firm covariance is constant over time

$$Cov(\varepsilon_{i,t}, \varepsilon_{j,t+k}) = \sigma_{ij} \text{ for all } t \text{ if } k=0$$
$$= 0 \text{ if } k \neq 0$$

from clustering standard errors by time

1. Firm's residuals can be autocorrelated only within-firm, not across firms

$$Cov(\varepsilon_{i,t}, \varepsilon_{i,t+k}) = \sigma_{ik} \text{ for all } t$$
$$Cov(\varepsilon_{i,t}, \varepsilon_{j,t+k}) = 0 \text{ for } i \neq j \ \& \ k \neq 0$$

③ Fixed Effect $y_{i,t} = \alpha_i + \beta x_{i,t} + \varepsilon_{i,t}$

Assume 2 firms ($N=2$) with firm fixed effects

$$Y = \begin{bmatrix} y_{1,1} \\ \vdots \\ y_{1,T} \\ y_{2,1} \\ \vdots \\ y_{2,T} \end{bmatrix} \quad X = \begin{bmatrix} 1 & 0 & x_{1,1} \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_{1,T} \\ 0 & 1 & x_{2,1} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_{2,T} \end{bmatrix} \quad \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} = (X'X)^{-1} X'Y$$

$$\beta = \frac{Cov(y_{i,t} - \bar{y}_i, x_{i,t} - \bar{x}_i)}{Var(x_{i,t} - \bar{x}_i)} \quad \text{where} \quad \bar{y}_i \equiv \frac{1}{T}\sum_{t=1}^{T} y_{i,t}, \quad \bar{x}_i \equiv \frac{1}{T}\sum_{t=1}^{T} x_{i,t}$$

Without fixed effect :

$$\beta = \frac{Cov(y_{i,t} - \bar{y}, x_{i,t} - \bar{x})}{Var(x_{i,t} - \bar{x})} \quad \text{where } \bar{y} \equiv \frac{1}{TN} \Sigma_{t=1}^{T} \Sigma_{i=1}^{N} y_{i,t}$$

Standard panel regression weights by the informativeness of each observation, whether it is across firms or time.

Fama-MacBeth : a kind of Panel approach,

Weight each time t coefficient the same when taking average

weight of years where only few firms = weight where many firms

14 factors : (12 industries dummies) + ln BM + ln Prof

$$X_t = \begin{bmatrix} 1 & lnProf_{1,t} & lnBM_{1,t} & indDum2_{1,t} & \cdots & indDum12_{1,t} \\ 1 & lnProf_{2,t} & lnBM_{2,t} & indDum2_{2,t} & \cdots & indDum12_{2,t} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & lnProf_{N_t,t} & lnBM_{N_t,t} & indDum2_{N_t,t} & \cdots & indDum12_{N_t,t} \end{bmatrix}$$

trade BM effect: We use the N_t Portfolio weights given by the 3^{rd} row of $(X_t' X_t)^{-1} X_t'$ $(14 \times N)$ where X_t

$$E[R_{Tot\ Port}] = E[R_{rf}] + k E[R_{Fama-MacBeth}]$$

$$\sigma[R_{Tot\ Port}] = k\,\sigma[R_{Fama-MacBeth}]$$

$$SR[R_{Tot\ Port}] = \frac{E[R_{Tot\ Port}] - E[R_{rf}]}{k\,\sigma[R_{Fama-MacBeth}]} = SR[R_{Fama-MacBeth}]$$

Simple Regression : (book to market)

1. direct effect – the return predictor & positive
2. "indirect" effect from industry and profitability (both correlated with b/m)

④ Omitted Variable Bias

true : $y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i$

we : $y_i = \alpha + \beta^* X_i + \varepsilon_i^*$

$$\Rightarrow \hat{\beta}_{SR} - \hat{\beta}_{MR} \approx \gamma \frac{Cov(X, Z)}{Var(Z)}$$

Solution: to vary X in an experimental fashion using random-ization. Random variation in X is not correlated with anything

Fama - MacBeth (multiple right-hand side variables)

Realized excess return on Portfolio k is the regression coefficient, $\lambda_{t,k}$

Panel Regression requires 假设: βs are constant over time & cross-sec

1. Be careful of using firm-level fixed effects because of small sample issues → average badly estimated. Use industry effect!
2. Year-fixed effects appropriate if interested in cross-sectional differences (difference of two stock's return on earnings)
3. Industry-fixed effects appropriate if each industry has permanent differences in the y variable (like RoE)
   - Many firms in each industry → fixed effect w/ little noise

⑤ Logistic Regression Model

$$Pr(Y=1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}$$

$V = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ as a score. $V\uparrow$, $Pr\uparrow$

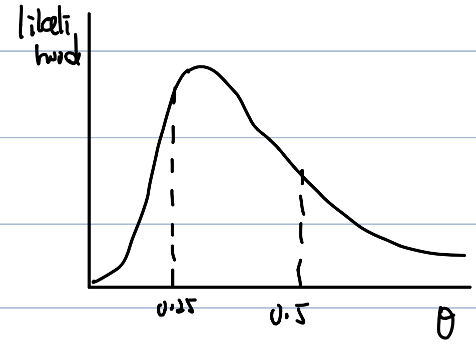
Not the same interpretation of logistic slope coefficients since it's no more linear.

$$\log(P/1-P) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

change in probability of $Y=1$ w.r.t X variable:

$$\frac{\partial Pr(Y=1|X)}{\partial X_j} = \beta_j Pr(Y=1|X)(1 - Pr(Y=1|X))$$

Likelihood: All probabilities for the observed defaults multiplied together.



N = 10 toss

head = 3 times

$L(\theta) = \theta^3 (1-\theta)^{10-3}$

$$L(\beta \mid y, X) = \prod_{i=1}^{N} Pr(Y_i = 1)^{y_i} \left(1 - Pr(Y_i = 1)\right)^{1-y_i}$$

y_i: 1 or 0

$$Pr(Y_i = 1) = f(\beta) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} \quad \text{where} \quad X = \begin{bmatrix} x_1' \\ \vdots \\ x_N' \end{bmatrix}$$

☆ Deviance: A measure of fit for a logistic regression (R^2)

Saturated model: M_s (enough parameters to fit each observation perfectly, so that $L = 1$)

null Model: M_n (natural lower bond)

the one with all coefficients $= 0$, except the intercept coefficient

Proposed model: M_c (candidate model with k betas all estimated)

☆ null deviance: $d_{null} = 2(\ln L(M_s) - \ln L(M_n))$

residual deviance: $d_{residual} = 2(\ln L(M_s) - \ln L(M_c))$

Analogue of SSE (smaller is better)

Standard likelihood ratio test, difference is Chi2-distributed with degrees of freedom = # of observations - # of parameters in the non-saturated models

null model for FICO: Only an intercept term and no slope.
A model the intercept is chosen to make the (fitted prob. that $Y=1$) = (the frequency for which $Y=1$) in the data.

☆ Inclusion - Exclusion Test (For removed factors)
Compare the deviance from the full (all variables) with the restricted (不重要的 variables removed)
how many variables thrown out   new\$df - old\$df
Change in fit: new\$dev - old\$dev   (+ means worse)

Lift Table: to evaluate ability of model to predict default
Sort the data by fitted probabilities and compute the mean of the Y variable (mean default rate) for each decile of fitted probabilities.
If good model → higher default rates for higher fitted prob.

Error Type

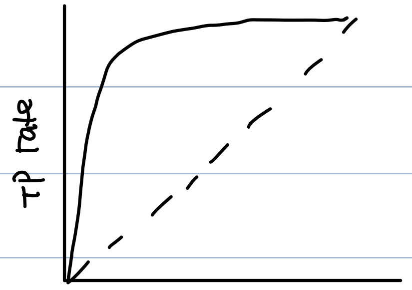
	True Default	True Paid
Predicted Default	2 (TP)	0 (FP)
Predicted Paid	1 (FN)	3 (TN)

Model predicting default:
True Positive = 2
False Positive = 0

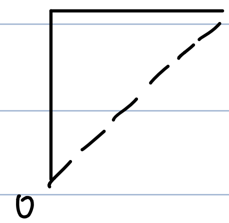
$FPR = FP / FP+TN$ $TPR = TP / TP+FN$

ROC curve (Receiver Operating Characteristics graphs)
Tracing out T/F Positives for different cutoffs yields the data for scatter

plot



ideal:



AUC (Area Under Curve): model fit in 1 number

probability of positive instance > negative

not directly related

Profit Maximization

$$\max_{\{cutoff\}} \quad \underset{\uparrow \text{ good}}{TN(cutoff)} \times Profit_{no\ Def.} - \underset{\uparrow \text{ bad}}{FN(cutoff)} \times Loss_{default}$$

↑ st λ

$TN = N - FP = N(1 - FPR)$ $\quad$ N: Total Negative (no defaults)

$FN = P - TP = P(1 - TPR)$ $\quad$ P: Total Positives (default)

⑥ Overfitting

In sample we have $E[f(X_{i,t})\, e_{i,t+1}] \neq 0$

$\Rightarrow E_t[R^e_{i,t+1}] \neq f(X_{i,t})$

↓

Shrinkage $\quad$ decrease parameters for variables. (better out of sample)

$$\mu_i^{shrink} = w_i \mu_i + (1 - w_i)\, \mu_{prior}$$

unconditional mean of prior distribution

Bayesian Inference: Use probability statements to assess our views about model parameters (like regression coef.)

Joint distribution: $P_{X,Y}(x, y) = P_{Y|X}(y \mid x)\, P_X(x)$

$$P_X(x) = \int P_{X,Y}(x, y)\, dy$$

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{matrix} \mu_X \\ \mu_Y \end{matrix}, \ \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

$X \sim N(\mu_X, 1)$

$Y | X = x \sim N(\mu_Y + \rho(x - \mu_X), 1 - \rho^2)$

stock returns i.i.d $\theta = (\mu, \sigma^2)$

A particular stock with T observations, the likelihood is

$$P(R_1^e, R_2^e, \dots, R_T^e \mid \mu, \sigma^2) = \prod_{t=1}^{T} P(R_t^e \mid \mu, \sigma^2)$$

$$= (2\pi\sigma^2)^{-T/2} \cdot \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{T} (R_t^e - \mu)^2\right)$$

$$P(\theta \mid D) \propto P(D \mid \theta) \cdot P(\theta)$$

PDF: $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$

⑦ Shrinkage Estimator

$$\mu_i^{shrunk} = w_i \frac{1}{T} \sum_{t=1}^{T} R_{i,t}^e + (1 - w_i)\mu_{prior}$$

where $w_i = \dfrac{T\sigma_i^{-2}}{T\sigma_i^{-2} + \sigma_{Prior}^{-2}}$ 　　 ↑ 既的 belief

Prior estimate from cross-section

$$\sigma_i^2(\mu_i) = \sigma_i^2(\hat{\mu}_i) - \frac{1}{N} \sum_{i=1}^{N} [\text{st. error}(\hat{\mu}_i)]^2$$

※ multiple regression: $\beta^{Post} = W\hat{\beta} + (I_k - W)\bar{\beta}$

where $W = ((X'X)\sigma^{-2} + A\sigma^{-2} I_k)^{-1}(X'X)\sigma^{-2}$

$$= (X'X + AI_k)^{-1}(X'X)$$

⑧ Ridge Regressions: Overfitting means regression coefficients (βs) are too big in absolute value

⇒ Penalize objective function when coefficients too big

$\Rightarrow$ shrink coefficients toward zero.

OLS objective function

OLS: $y_i = \beta_0 + \sum_{k=1}^{K} \beta_k X_{ki} + \epsilon_i$

objective f(.): $\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{k=1}^{K} \beta_k X_{ki})^2$

Ridge regression objective function L_2

$$\min_{\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{k=1}^{K} \beta_k X_{ki})^2 \quad \text{s.t.} \quad \sum_{k=1}^{K} \beta_k^2 \le B$$

$$\min_{\beta} \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta x_i)^2 \quad \text{s.t.} \quad \beta'\beta \le B$$

$$\min_{\beta} \frac{1}{2} \{ (Y - X\beta)'(Y - X\beta) + \lambda \beta'\beta \}$$

$$-X'(Y - X\beta) + \lambda \beta = 0 \Rightarrow \beta = (X'X + \lambda I_K)^{-1} X'Y$$

B: "the coefficient budget" the regression is given.

Un-demean all variables (y and X) by setting $\hat{\beta}_0 = \bar{y} - \sum_{k=1}^{K} \hat{\beta}_k \bar{X}_k$

in Penalty form: $\min_{\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{k=1}^{K} \beta_k X_{ki})^2 + \lambda \sum_{k=1}^{K} \beta_k^2$

B and $\lambda > 0 \Rightarrow$ one-to-one mapping

(9) Cross-validation: out-of-sample model selection criterion

- split up the sample in random training and test sets & estimate performance on test sets.
- Choose the level of the constraint that gives best prediction.

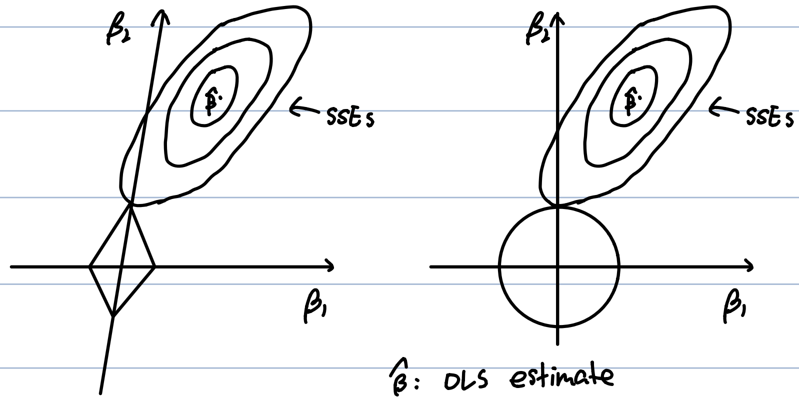
K-folds:

1. Split sample in K equal-sized groups (typically 5/10)
2. For each of the K fold, model on the other K-1 folds & test on the Kth fold. (K out-of-sample tests)

3. Prediction error is the basis for choosing best model.

The LASSO $\alpha = 1$ for glmnet ↗ 0 for Ridge

$$\min_{\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{k=1}^{K} \beta_k X_{ki})^2 \text{ s.t. } \sum_{k=1}^{K} |\beta_k| \le B$$



absolute value constraint makes corner solution with 0 coefficients much more likely than in OLS or Ridge.

Elastic net $(0 < \alpha < 1)$ (highly correlated X values w/ LASSO)

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \sum_{k=1}^{K} \beta_k X_{ki})^2 \text{ s.t. } \sum_{k=1}^{K} (1-\alpha)\beta_k^2 + \alpha|\beta_k| \le B$$

⑩ Bayes Regression: $y | X \sim N(X\beta, \sigma^2 I_n)$

$$P(y_1, \cdots, y_n | \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i'\beta)^2\right)$$

need prior distribution.

Bayes Estimator: Posterior mean

Prior "precision" matrix $A \uparrow \rightarrow \bar{\beta}$ prior ↓

$$\tilde{\beta} = (X'X + A)^{-1}(X'y + A\bar{\beta}) \quad \text{even co-linearity}$$

$$\tilde{\beta} = (X'X + A)^{-1}(X'X\hat{\beta} + A\bar{\beta})$$

if $X'X$ is large (large N / lot of variation in data)

→ little shrinkage

If one variable has only a tiny bit of variation

→ its coefficient will get shrunk a lot while others not

Bayesian interpretation of Ridge / Lasso allows for

Computation of Posterior distribution of the regression coefficients.

$\Rightarrow$ Allow for Statistical inference and tests

Just Ridge / Lasso 本身 不能 用 standard OLS error apparatus

⑪ MVE Portfolio

$$R^e_{MVE,t+1} = \sum_{i=1}^{N_t} W_{i,t} R^e_{i,t+1}$$

stock characteristics : $X_{i,t}$ — $k \times 1$ vector for each i & t $\sim N(0,1)$

$W_{i,t} = b' X_{i,t}$ b : $k \times 1$ Constant Vector

$$R^e_{MVE,t+1} = \sum_{i=1}^{N_t} b' X_{i,t} R^e_{i,t+1} = b' \sum_{i=1}^{N_t} X_{i,t} R^e_{i,t+1} = b' F_{t+1}$$

F_{t+1} : $k \times 1$ vector of factor returns. from $X_{i,t}$

MVE: $b = \Sigma^{-1} E[F_{t+1}]$ Σ : var-Cov matrix of factor returns.

$\Sigma b = E[F_{t+1}]$. $\bar{F}_j = \sum_{k=1}^{K} b_k \, Cov_T(F_{k,t}, F_{j,t}) + \varepsilon_j$

⑫ Machine Learning

Supervised Learning : both response variable (y) and predictors (x) available. (Right answer available)

Unsupervised Learning : discover interesting things about the measurements of the P "features" $X_1, X_2, \cdots, X_p$

1. PCA : representative variables explaining most of the variation
2. Clustering : Partition data into similar groups

Fundamental trade-off (Bias - Variance)

$MSE = E[(Y - \hat{f}(x))^2]$, $Y = f(x) + \varepsilon$

$$= E[(f(x) - \hat{f}(x))^2] + Var(\varepsilon)$$

$$= \underbrace{Var(\hat{f}(x))}_{\text{standard error}^2} + \underbrace{E[f(x) - \hat{f}(x)]^2}_{\text{bias}^2} + Var(\varepsilon)$$

$$= Var(\hat{f}(x)) + [Bias(\hat{f}(x))]^2 + Var(\varepsilon)$$

In general, more flexible method have higher variance but result in less bias.

Regularization (Shrinkage) introduces bias but reduces variance (shrinking coefficients towards zero.

Cross-Validation: attempt to find the optimal trade-off between variance and bias

⑬ Decision Tree

J: the fixed number of terminal nodes / boxes / leaves.

$\hat{y}_{R_j}$: the average of the observations in box R_j

objective function: minimize the sum of squared errors.

$$\min_{\{R_j\}_j} \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

$J=2$, N observations: N possible splits

$J=3$: order of N^2 possible splits

Complexity of problem grows exponentially

Recursive binary splitting (No look ahead for global optimal)

1. Select one predictor, X_j
2. Find the one split, (2 boxes) min. the sum of squares.

save the breakpoint $X_j = s$ for each j

3. Loop through all predictors and choose the one leading to lowest SSE out of all predictors. ⇒ The first breakpoint
4. For 每个 predictor, find the split min. the SSE (any box).
5. keep going until a convergence criterion is met

No region > 10 observations / # of end nodes = 30

overfitting of each 单独 tree

Bagging (Bootstrap aggregation)

Averaging reduces variance for independent observations

Use bootstrap, taking B repeated samples from original dataset and fit B trees.

$\frac{2}{3}$ data each tree

$\frac{1}{3}$ for cross-validation

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{b}(x)$$

Random Forest: de-correlate the samples from bootstrap.
Random selection of m of the p predictors are chosen for each bootstrapped sample. ⇒ Reduce correlation due to less overlap.
Typically $m = \sqrt{p}$
In the end average the prediction from all trees. (Like bagging)

(14) Boosting: fits small trees and learns slowly by adding small trees fit to the prediction errors of the existing trees.

Tuning Parameters: 1. Number of trees (B / γ)
2. Shrinkage parameter λ Small λ → large B to fit data
3. number of splits in each tree d- interaction depth

$d = 1$ → fitting an additive model (no interactions)

Algorithm: 1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for $\forall i$ in training set

2. For $b = 1, 2, \cdots, B$, repeat

a) Fit a tree $\hat{f}^b$ with d splits ($d+1$ terminal nodes) to the training data (X, r)

b) Update $\hat{f}$ by adding in a shrunken version of the new tree

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

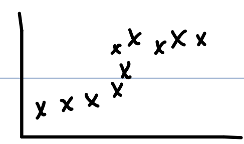
c) Update the residuals $\quad r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$

3. Output the boosted model

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

minimize residuals one step at a time

XGBoost eXtreme Gradient Boosting



$\Omega(f)$ the Complexity of tree

$L(f)$ the prediction error loss function (variance)

$$\min Obj(\Theta) = \min (L(\Theta) + \Omega(\Theta))$$

Common loss function for tree $\quad L = \Sigma_i (y_i - \hat{y}_i)^2$

Common regularization term $\quad \Omega(f) = \gamma N + \frac{1}{2}\lambda \Sigma_{n=1}^{N} \beta_n^2$

N: number of leaves (boxes, nodes), f: prediction function

$$f = \Sigma_{n=1}^{N} \beta_n \cdot 1_{(x \in R_n)}$$

Like ridge but additional Penalty for size of tree (N)

Run usual FMB Regressions to get Portfolio Performance based on sample

XGBoost better than Random Forest for out of sample performance.

PCA: symmetric $m \times m$ matrix B has a spectral decomposition

$$B = P \Lambda P'$$

Λ: diagonal matrix with eigenvalues λ on diagonal > 0

$P^T = P^{-1}$ $\quad P$: $m \times m$ orthogonal matrix consisting of m eigenvectors.

When apply PCA to Σ, $(N \times N)$ each eigenvector defines the portfolio weights in a portfolio with variance = eigenvalue.

The factor explains most σ^2 (most covariance between stocks; factor 1) is $F_{1,t} = \sum_{n=1}^{N} P(n,1) R^e_{n,t}$

$P(n,1)$: n'th row in column 1 of matrix P

all factors from the PCA are uncorrelated.

APCA: Asymptotic Principle Component Analysis

good when N大 T小

Assume returns follow a k-factor model

1. Take a relevant sample of stock returns (perhaps last year of daily data)
2. Let Rt denote the T by N matrix of stock returns in this sample
3. Let Ω = Rt * Rt'
4. Get eigenvectors and eigenvalues of Ω
5. The eigenvectors corresponding to the *K* largest eigenvalues are the *T* returns to the *K* factors of the underlying factor model (up to a constant of proportionality).

PCA λ : weight

Use factors to hedge movements in asset values.

1. Run Fama-MacBeth at each t including both signal and factor: $R_{it} = \beta_{i1} F_{1t} + \beta_{i2} F_{2t} + \varepsilon_{it}$

F: variance 最大

Regress R_{it} on F_{1t} & F_{2t} $\Rightarrow \beta_{i1}$ & β_{i2}

2. $R_{it} = \lambda_{0t} + \lambda_{1t} \beta_{i,t-1} + \lambda_{2t} \underbrace{X_{i,t-1}}_{\text{your signal}} + \varepsilon_{it}$

15 Non-linear clustering techniques: K-means clustering

Partition data into K distinct, non-overlapping clusters.

$C_1, C_2, \dots, C_K$: sets containing indices in each cluster

1. $C_1 \cup C_2 \cup \dots \cup C_K = \{1, 2, \dots, n\}$ 有且仅有一个

2. $C_i \cap C_j = \phi$ for $i \neq j$

$f(\cdot)$: $\min_{C_1, \cdots C_K} \sum_{k=1}^{K} W(C_k)$ $\quad$ $W(C_k)$: within-cluster variation of cluster k

Squared Euclidean distance: $W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$

$|C_k|$: # of data in k

⑯ Unstructured

remove Punctuation ("." ";" "?" ",") $\quad$ tm_map

Stop words ("English") 介词 "a", "the", "it" etc

stemDocument: Stemming (invest kept for invest, investing, invested ...)

stripWhitespace: only seperated by one space (remove extra)

Corpus_matrix: Contains all words along with Count

Corpus: full set of text data

Latent Dirichlet Allocation (LDA): extract text topics

LDA: 1. Decide number of words N in document D

2. Choose a topic mixture. ($\frac{1}{3}$ food + $\frac{2}{3}$ animals for D)

3. Generate each word w_i in the document

first pick a topic, then use it to generate the word

4. Use the documents backtrack to find topics

document_sums: $K \times D$ matrix, # of times words in each document (column) were assigned to each topic (row)

topics: # a word (column) was assigned to a topic (row) topics model : a simple measure of similarity between 2 documents in terms of their themes.

assume that the number of words contained within both topics 1 and 2 is similar to the length of each of the documents. Armed with the assumption, I propose the following measure:

$$Pr(\text{topic } 1_i) \cdot Pr(\text{topic } 1_j) + Pr(\text{topic } 2_i)\, Pr(\text{topic } 2_j)$$

Higher fees translate into lower alpha, as expected, given that management fees eat into fund's performance. Also, bigger funds (higher NAV) underperform relative to smaller funds (lower NAV). The second fact is a well-known puzzle in the finance academic literature.