FMB: Run T cross-sectional Regressions.

$R^e_{i,t} = \lambda_{0,t} + \lambda_{1,t} \ln BM_{i,t-1} + \varepsilon_{i,t}$. save all the T estimated $\lambda_{1,t}$ and obtain $\hat{\lambda}_1 = \frac{1}{T}\sum_{t=1}^{T} \lambda_{1,t}$. $Var(\hat{\lambda}_1) = Var_T(\lambda_{1,t})/T$

$$\lambda_{1,t} = \frac{1}{N_t} \underbrace{\frac{\ln BM'_{t-1} - E_i(\ln BM_{i,t-1})}{Var_i(\ln BM_{i,t-1})}}_{1\times N} \cdot \underbrace{R^e_t}_{N\times 1}$$

across stock at a time

Weight of FMB ($\sum w_{i,t} = 0$) $w_{i,t-1} = \frac{1}{N} \frac{x_{i,t-1} - E_i[x_{i,t-1}]}{Var_i[x_{i,t-1}]}$

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

t-stat $t = \left(\frac{\hat{\lambda}_1}{\sqrt{Var_T(\lambda_{1,t})}}\right)\sqrt{T}$ ← SR ratio

FMB Regression test is a test of whether a traded long-short portfolio has a high SR.

Panel regression: Two dimensions. typically cross-section and time series.

Balanced panel: N observations in cross-section for each t.

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

$cov(X,Y) = E([X - E[X]]\cdot[Y - E[Y]]) = E(XY) - E(X)\cdot E[Y] = \frac{1}{N}\sum_{i=1}^{n}(X_i - E[X])(Y_i - E[Y])$.

Implicit Assumption: slope coefficients do not vary over time or cross firms

Intercept can be varied: over time or across firms.

Canonical Panel Regression $y_{i,t} = \delta_t + \theta_i + \beta' x_{i,t} + \varepsilon_{i,t}$ → effect is the same.

δ_t: time-fixed effect; θ_i: firm fixed effect

Clustering of Panel Regression.

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

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

from clustering standard errors by time.

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

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

Fixed effect: $y_{i,t} = \theta_i + \beta x_{i,t} + \varepsilon_{i,t}$

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

$$Y = \begin{bmatrix} y_{1,1} \\ \vdots \\ y_{1,T} \\ y_{2,1} \\ \vdots \\ y_{2,T} \end{bmatrix} \quad X = \begin{bmatrix} 1 & 0 & x_{1,1} \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_{1,T} \\ 0 & 1 & x_{2,1} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_{2,T} \end{bmatrix} \quad \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} = (X'X)^{-1}\cdot X'\cdot Y \quad \beta = \frac{cov(y_{i,t} - \bar{y}_i,\ x_{i,t} - \bar{x}_i)}{Var(x_{i,t} - \bar{x}_i)}$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{i,t}$ $\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{i,t}$

without fixed effect: $\beta = \frac{cov(y_{i,t} - \bar{y},\ x_{i,t} - \bar{x})}{Var(x_{i,t} - \bar{x})}$ where $\bar{y} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} y_{i,t}$

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

FMB 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 3rd row of $(X_t' \cdot X_t)^{-1} X_t'$ ($14 \times N$) where X_t is above.

$E(R_{total\ port}) = E(R_{rf}) + kE[R_{FMB}]$ $\quad \sigma[R_{total\ port}] = k\sigma[R_{FMB}]$

$$SR[R_{total\ port}] = \frac{E[R_{total\ port}] - E[R_{rf}]}{k \cdot \sigma[R_{FMB}]} = SR[R_{FMB}]$$

Simple Regression (b-to-M) 1. direct effect - the return predictor & Positive

2 indirect effect from industry and profitability (both correlated with B/M)

Omitted Variable Bias.

true: $y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i \Rightarrow \hat{\beta}_{SR} - \hat{\beta}_{MR} \approx \gamma \cdot \frac{cov(x,z)}{Var(z)}$

we $y_i = \alpha + \beta^* x_i + \varepsilon_i^*$

Solution: to vary X in an experimental fashion using randomization.

Random variation in X is not correlated with anything.

FMB. (Multiple RHS variables) Realized excess return on portfolio K is the regression coefficient. $\lambda_{t,K}$. Panel Regression requires assumption.

βs are constant overtime & cross-sec

1. Be careful of using firm-level-fixed effects because of small samples issues → average badly estimated. Use industry effect

2. Year-fixed effects. appropriate if interested in cross-sectional differences (difference of two stocks 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 with little noise

Logistic Regression Model.

$$Pr(Y=1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K)}{1 + \exp(\beta_0 + \beta_1 X + \cdots + \beta_K X_K)}$$

$v = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K$ as a score. $v \uparrow$ $P_r \uparrow$

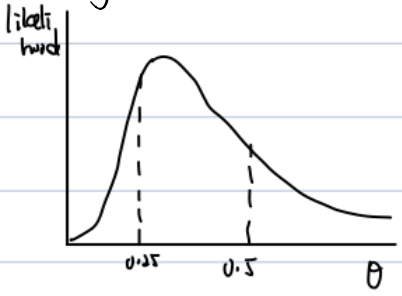
Not the same interpretation of logistic slope coefficients.

since it's no more linear. $\log(p/1-p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K$

Changing 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

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

$$L(\beta | y, X) = \prod_{i=1}^{N} Pr(Y_i = 1)^{y_i} (1 - Pr(Y_i = 1))^{1-y_i}$$ ($1-y_i$: 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 bound)

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 chi 2 distributed with degrees of freedom = # of observations — # of parameters in the non-saturated models.

a) Give the Fama-MacBeth regressions you will run. Use clear notation. (6 pts)

Run the cross-sectional regression $R^e_{i,t} = \lambda_{0,t} + \lambda_{1,t} z_{i,t-1} + \varepsilon_{i,t}$ across stocks $i = 1, ..., N$ at each time t, for a total of T cross-sectional regressions.

b) Using matrix algebra, give an expression for the portfolio weights at each time t for the portfolio that replicates the estimated Fama-Macbeth coefficient on the trading signal from the previous question. Take care the define any matrices used. (6 pts)

The portfolio weights are (from the class notes):

$$w_{i,t} = \frac{1}{N} \frac{z_{i,t} - E_N[z_{i,t}]}{var_N(z_{i,t})}$$

where the subscript N denotes a moment taken across stocks (cross-sectional mean and variance.

The Fama-MacBeth regression coefficient of interest ($\lambda_{1,t+1}$) is then:

$$\lambda_{1,t+1} = \sum_{i=1}^{N} w_{i,t} R^e_{i,t+1}, \qquad \text{which is a portfolio return.}$$

d) Explain why adding industry dummies might improve the Sharpe ratios of your implied trading strategy. (6 pts)

Here, you simply include firms' book-to-market ratios in each Fama-MacBeth regression:

$$R^e_{i,t} = \lambda_{0,t} + \lambda_{1,t} z_{i,t-1} + \lambda_{2,t} bm_{i,t-1} + \varepsilon_{i,t}$$

We are still interested in $\hat{\lambda}_1$, defined as in the previous sub-question. The regression now effectively asks, what is the Sharpe ratio of a strategy based on the signal, holding firm book-to-market ratios constant. Let's take a simple example where the signal is book-to-market plus a little noise. Now, (ignoring estimation error) we would get $\lambda_1 = 0$, as holding book to market constant the signal would be a sort based only on noise. In the univariate regressions, however, the signal-based sort would be a sort on book-to-market (as well as a little noise) which would likely yield some excess return. In sum, we are estimating the marginal effect of the signal.

i. Explain clearly what an Elastic Net constraint is. (5 pts)

The constraint is an L_1-constraint, which means the we put a constraint on the sum of the absolute values of each coefficient. This is compared to an L_2-constraint, where the constraint is on the sum of squared coefficient values, or OLS which has no constraint. The absolute value constraint produces kinks that make corner solutions with zero coefficients much more likely than in OLS or Ridge regressions, where the likelihood of an

exactly zero coefficient is in fact zero. Also, recall the plot from the class notes for a visual depiction of this:

ii. Explain the K-fold cross-validation procedure often used to find the 'optimal' value of the constraint parameter 'lambda.' (7 pts)

Divide the data into K equal subsamples. Then, estimate the model based on K-1 subsamples and find mean squared error of model prediction on the K'th out-of-sample fold. Do this for all K-1 permutations. Choose the lambda that minimizes the average mean squared error across the K out of sample folds.

i. What an ROC curve is. In particular, what goes on each axis and how do you define these variables.

We use ROC curves to compare the performance of predictive models. An ROC has the true positive rate (also called sensitivity) on the y-axis and the false positive rate (100-specificity) on the x-axis. The true positive rate represents the proportion of observations with a certain characteristic (think of days during which the stock market is up) that are correctly identified to possess that characteristic. The false positive rate is the proportion of observations that do not possess the characteristic that are erroneously determined to possess the characteristic (think of days during which the stock market is down but our model states that the stock market is up).

Based on the scenario described in this part, we would conclude that our model works fine for some observations but fails for others. An example of such a scenario is the value effect related to small and big stocks. A reasonable approach would be to only use our predictive model for cases in which $x < 0.5$.

(i) Corpus is all the text documents you are working with.

(ii) Stopwords are words that are very common (e.g., "and", "as", "the", etc) that we want to remove from the text as they are very uninformative.

(iii) Stemming means cut words down to their 'stem' their shortest core. For instance, "invests" and "invested" both have the stemmed form "invest." This is done so the computer counts these words as the same and not different.

Both size and management fees predict firm performance. 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. Note that in our case NAV and fee breakpoints provide the same information.

c) Briefly explain how *boosting* works to improve the mean squared error of the decision tree's prediction error. (6 pts)

With bagging, you resample, typically with replacement, and create J samples from your data set (really, a bootstrap procedure). Estimate J trees based on these samples. Get the prediction for each tree and average predictions to arrive at your final prediction.

Each tree is estimated with noise and as long as this noise is not perfectly correlated across trees, there is value in taking an average to reduce noise and thereby improve the signal to noise ratio of you predictions.

(d) Briefly explain the main difference between linear regression models and decision trees. Use equations to illustrate your logic.

The two models also differ in functional form assumptions:

- Linear regression:

$$f(X) = \beta_0 + \sum_{j=1}^{P} X_j \beta_j$$

- Decisions trees:

$$f(X) = \sum_{m=1}^{M} \beta_m \mathbb{I}\{X \in R_m$$

The intuition is that with decision trees dummy variables allow for non-linear 'buckets' and the definition of the dummy variables is endogenous to the procedure.

Both approaches use the same optimization logic to obtain the coefficients of interest: we minimize the sum of squared errors. However, the minimization problems take on a different form:

- Linear regression :

$$min_{\tilde{\beta}} \left\{ \left(y - X\tilde{\beta}\right)' \left(y - X\tilde{\beta}\right) \right\}$$

- Decision trees:

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

μ_i is a firm fixed effect. We include fixed effects to remove the "between" variation (between firms) and focus on "within" variation.

$$E_t\left[rv_{i,t+1} + rv_{i,t+2}\right] = 2\mu_i + \mu_i\delta_1 + \delta_1 rv_{i,t} + \delta_1^2 rv_{i,t}$$

Our signal is the spread: $E_t\left[rv_{i,t+1} + rv_{i,t+2}\right] - iv_{i,t}^2$. I call the first term RV and the second term IV for simplicity. IV and RV are not our signal individually. Option prices increase in the volatility of the underlying. Therefore, if $IV > RV$, the option is overpriced, i.e. we would like to short the option. Conversely, if $IV < RV$, we would like to go long in the option.

Each month we compute the spread and form portfolios based on the magnitude of the spread.

$$min_\beta \sum_{t=1}^{T}(y_{t+1} - \beta x_t)^2 \qquad \beta^2 \leq B$$

Form the Lagrangian:

$$\mathfrak{L} = (Y - X'\beta)(Y - X'\beta) + \lambda(B - \beta^2)$$

Simply take the FOC w.r.t. β and obtaint the following expression for β:

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

In a ridge regression setting, we interested in obtaining a value for the λ parameter one of the approaches that provides us with an "optimal" value for $\lambda$is K-fold cross validation. Here is a quick description of how the procedure works:

We divide the sample into K subsamples and run a ridge regression separately for each group of K-1 folds. We use the Kth fold as an out-of-sample (OOS) test. Based on the OOS tests, we select the λ that minimizes the value of the MSE (you can also use SSE or any other appropriate decision criterion).

Cross-validation affects the β parameter through the fact that the procedure allows to choose the "optimal" λ.

All you need to do is to set the "budget" $B = 1$. The FOC is unchanged but the λ will change, which will affect the β.

LASSO uses an absolute value constraint with the following objective function:

$$min_\beta \sum_{t=1}^{T}(y_{t+1} - \beta x_t)^2 \qquad |\beta| \leq B$$

The LASSO constraint produces constraints that make corner solutions with zero coefficients more likely compared to a ridge regression (the likelihood of a zero coefficient with ridge is 0,

LASSO is better if we want to reduce dimensionality. Ridge is better if there is high correlation between the dependent variables.

Define a binary variable $y_{i,t} = \begin{cases} 1 & if\ rv_{i,t} > 0.01 \\ 0 & otherwise \end{cases}$

Based on this definition the likelihood function is (note that we have two dimensions, which most people ignored)

$$L(\beta) = \prod_{i=1}^{N}\prod_{t=1}^{T} p^{y_{i,t}}(1-p)^{1-y_{i,t}}$$

The **log likelihood** function is:

$$l(\beta) = \sum_{i=1}^{N}\sum_{t=1}^{T}\{y_{i,t}log(p) + (1 - y_{i,t})log(1-p)\}$$

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.

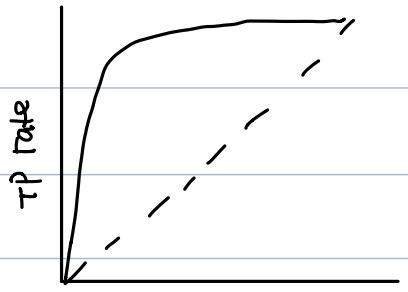
If good model → higher default rates for higher fitted prob.

Error Type	True Default	True Paid	Model predicting default:
Predicted Default	2 (TP)	0 (FP)	True Positive = 2
Predicted Paid	1 (FN)	3 (TN)	False Positive = 0

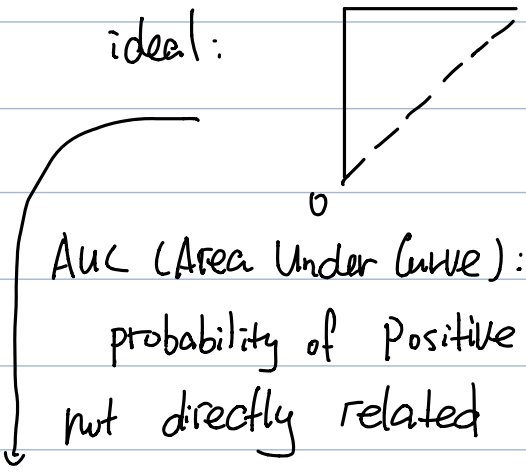
$FPR = FP/(FP+TN)$ $TPR = TP/(TP+FN)$

ROC curve (Receiver Operating Characteristics graphs)

Tracing out T/F Positives for different cutoffs yields the data for scatter



ideal:



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

probability of positive instance > negative

not directly related

Profit Maximization

$$\max_{\{Cutoff\}} \; TN(Cutoff) \times Profit_{No\ Def.} - FN(Cutoff) \times Loss_{default}$$

TN: good, Profit: fit ↗, FN: bad

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

$FN = P - TP = P(1 - TPR)$ — 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 — decrease parameters for variables. (better out of sample)

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

μ_{prior}: unconditional mean of prior distribution

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

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

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

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

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

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

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

A particular stock with T observations, the likelihood is

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

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

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

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 $w_i = \frac{T \sigma_i^{-2}}{T\sigma_i^{-2} + \sigma_{Prior}^{-2}}$ ↑ 貝氏的 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 $w = (X'X)\sigma^{-2} + A\sigma^{-2} I_k)^{-1} (X'X)\sigma^{-2}$

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

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

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

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

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

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

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.

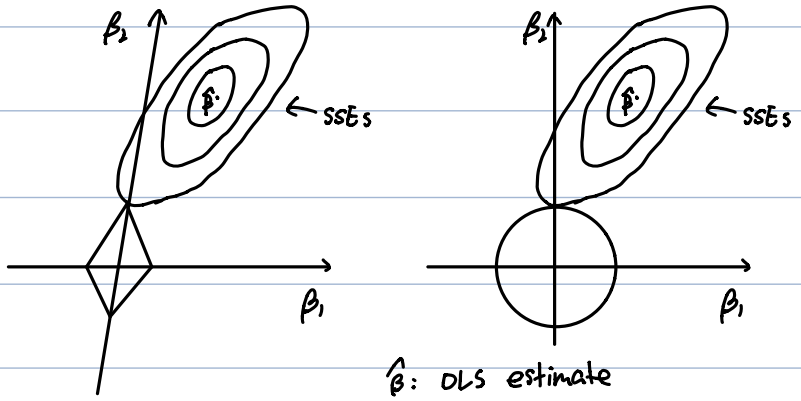
k-folds:

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)

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



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)

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

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

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

need prior distribution.

Bayes Estimator: Posterior mean (Prior "precision" matrix A: ↑ → 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.

⇒ 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 $\sim N(0,1)$

$w_{i,t} = b' X_{i,t}$ $\quad b$: $k\times 1$ 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: $b = \Sigma^{-1} E[F_{t+1}]$ $\quad \Sigma$: var-cov matrix of factor returns.

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

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 expohentially

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 单个 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

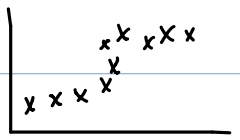
$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$

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

3. Output the boosted model $\quad$ minimize residuals

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \quad \text{one step at a time}$$

XG Boost 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 $\quad L = \sum_i (y_i - \hat{y}_i)^2$

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

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

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

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

Run usual FMB Regressions to get Portfolio Performance based on sample

XG Boost 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 $N \neq T$ 以上

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 量

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.

$C_1, C_2, \cdots, C_k$: sets containing indices in each cluster

1. $C_1 \cup C_2 \cup \cdots \cup C_k = \{1, 2, \cdots, n\}$ 2 在 2 的有一个

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

$f(\cdot)$: $\min_{C_1, \cdots C_k} \sum_{k=1}^{K} W(C_k)$ $W(C_k)$: 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

removePunctuation (=".", ";", "?", ":") tm_map

stopwords ("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 x 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)