Chapter7 Linear Models

From Risk Factors to Return Forecasts

Yanrui Zhou Wenchu Liu Shuxian Mei

Fintech Stu
Wuhan University

November10,2022

Linear Models

Multiple linear regression

The family of linear models represents one of the most useful hypothesis classes. Many learning algorithms that are widely applied in algorithmic trading rely on linear predictors.

- **efficiently trained**
- **robust to noisy financial data**
- **easy to interpret**
- **provide a good baseline**

How to formulate the linear model

In the population, the linear regression model has the following form for a single instance of the output y an input vector $X^T = [x_1, , x_p]$ and the error term ϵ:

$$y = f(x) + \epsilon = \beta_0 + \beta_1 x_1 \ldots + \beta_p x_p + \epsilon = \beta_0 + \sum_{j=1} \beta_j x_j + \epsilon$$

compactly, we have the matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

Ordinary least squares

To minimize the sum of the squared distances between the output value and the value generated by the linear model.

$$\begin{aligned} \mathrm{RSS}(\beta) &= \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta) \end{aligned}$$

Thus, the least-squares coefficients β^{LS} are computed as:

$$\underset{\beta^{LS}}{argmin}\ RSS(\beta) = (y - x\beta)^T (y - X\beta)$$

Ordinary least squares

$$\boldsymbol{y} = N \times 1 \qquad \left(\boldsymbol{y}^\top - \beta^\top \boldsymbol{X}^\top\right)(\boldsymbol{y} - \boldsymbol{X}\beta)$$
$$\boldsymbol{X} = N \times p \qquad = \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\beta - \beta^\top \boldsymbol{X}^\top \boldsymbol{y} + \beta^\top \boldsymbol{X}^\top \boldsymbol{X}\beta$$
$$\beta = p \times 1 \qquad = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \boldsymbol{X}\beta + \beta^\top \boldsymbol{X}^\top \boldsymbol{X}\beta$$

From two conclusions of matrix derivation:

$$\frac{\partial x^\top A x}{\partial x} = \left(A + A^\top\right) x \qquad \frac{\partial A^\top x}{\partial x} = A$$

After derivation, we get:

$$-2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\beta = 0$$
$$\widehat{\beta} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Ordinary least squares

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} = \widehat{\boldsymbol{y}} + \boldsymbol{\epsilon}$$

Geometric interpretation:
the coefficients that minimise RSS ensure that the vector of residuals $\boldsymbol{y} - \widehat{\boldsymbol{y}}$ is orthogonal to the subspace of R^P spanned by the P columns of $\boldsymbol{X}$, and the estimates $\widehat{\boldsymbol{y}}$ are orthogonal projections into that subspace.

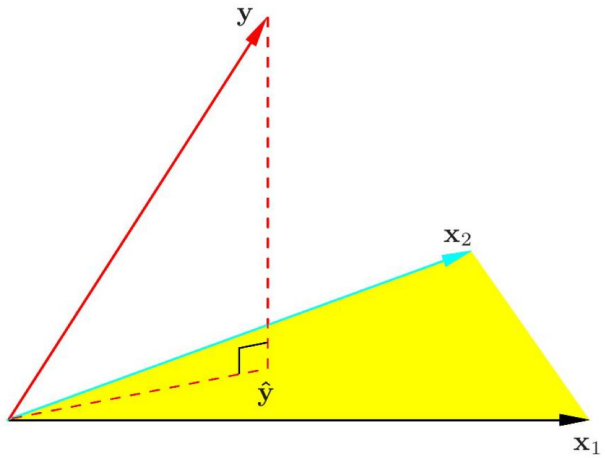


Figure: $\boldsymbol{y} = (\boldsymbol{x}_1, \boldsymbol{x}_2)\beta + \boldsymbol{\epsilon}$

Ordinary least squares

$$prove: \|y - \hat{y}\| < \|y - v\|$$

$y - \hat{y}$ is orthogonal to plane W:

$$(y - \hat{y})(\hat{y} - v) = 0$$

$$\because \quad y - v = (y - \hat{y}) + (\hat{y} - v)$$

$$\therefore \|y - v\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - v\|^2$$

$$\because \quad \|\hat{y} - v\| \neq 0$$

$$\therefore \quad \|y - v\|^2 - \|y - \hat{y}\|^2 > 0$$

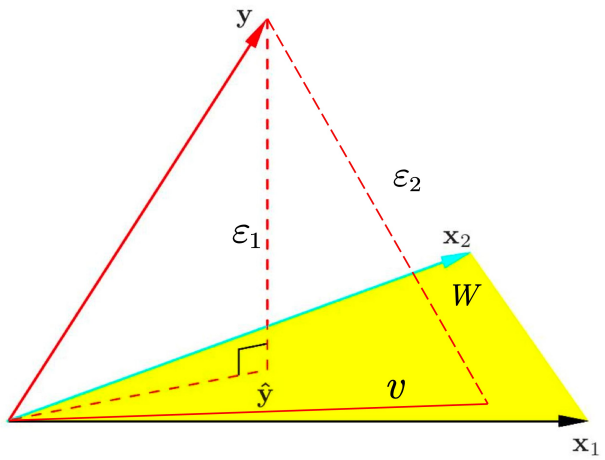


Figure: OLS prove

Maximum likelihood estimation

likelihood function: how likely it is to observe the sample of outputs when given the input as a function of the model parameters.
Let's set up the likelihood function by assuming a distribution for the error term, such as the standard normal distribution:

$$\epsilon_i \sim N(0,1)$$

since normal distribution:$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
the conditional probability of observing a given output y_i:

$$p\left(y_i \mid x_i, \beta\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(y_i - x_i\beta\right)^2}{2\sigma^2}}$$

Maximum likelihood estimation

Assuming the output values are conditionally independent, given the inputs, we have:

$$L(y, x, \beta) = \prod_{i=1}^{N} p(y_i \mid x_i, \beta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^{N}(y_i - x_i\beta)^2}{2\sigma^2}}$$

$$\log L(y, x, \beta) = n \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{\sum_{i=1}^{N}(y_i - x_i\beta)^2}{2\sigma^2}$$

Gradient descent

Gradient descent is a general-purpose optimization algorithm that will find stationary points of smooth functions.gradient?

$$\left.\frac{\partial f}{\partial l}\right|_{(x_0,y_0)} = \lim_{\Delta s \to 0} \frac{f\left(x_0 + \Delta s \cos\alpha, y_0 + \Delta s \cos\beta\right) - f\left(x_0, y_0\right)}{\Delta s}$$

$$\overset{differentiable}{=} \lim_{\Delta s \to 0} \frac{f_{x_0}' \Delta s \cos\alpha + f_{y_0}' \Delta s \cos\beta + o(\Delta s)}{\Delta s}$$

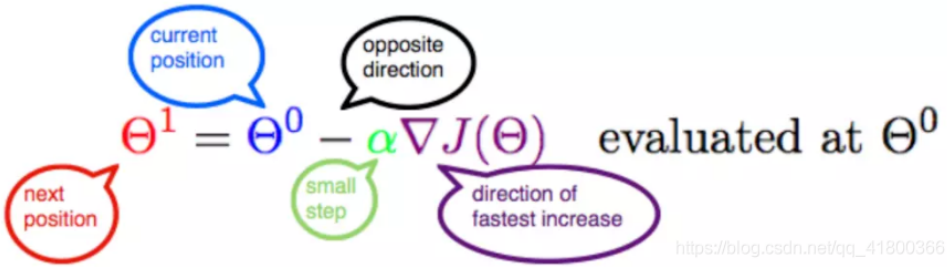
$$\because \nabla f\left(x_0, y_0\right) = f_{x_0}' i + f_{y_0}' j \qquad (definition\ of\ \ Gradient)$$

$$\therefore \left.\frac{\partial f}{\partial l}\right|_{(x_0,y_0)} = \nabla f\left(x_0, y_0\right) \cdot e_l = f_{x_0}' \cos\alpha + f_{x_0}' \cos\beta$$

$$= \left|\nabla f\left(x_0, y_0\right)\right| \left|e_l\right| \cos\theta$$

It turns out that the maximal change of the function value results from a step in the direction of the gradient itself.

Gradient descent



minimise the cost of a prediction error:

$$J(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)^2$$

$$h_\Theta \left(x^{(i)} \right) = \Theta_0 + \Theta_1 x_1^{(i)}$$

$$\nabla J(\Theta) = \left\langle \frac{\delta J}{\delta \Theta_0}, \frac{\delta J}{\delta \Theta_1} \right\rangle$$

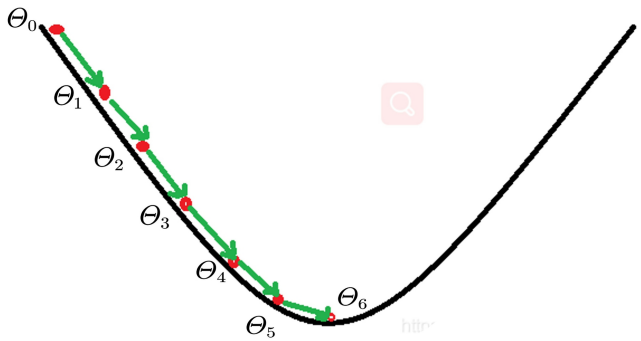


Figure: the process of GD

Gradient descent

For example, we have: $\mathrm{J}(\Theta) = \theta_1^2 + \theta_2^2 \quad \Theta^0 = (1,3) \quad \alpha = 0.1$

then, $\nabla \mathrm{J}(\Theta) = \langle 2\theta_1, 2\theta_2 \rangle$

$\Theta^0 = (1,3)$

$\Theta^1 = \Theta^0 - \alpha \nabla \mathrm{J}(\Theta) = (1,3) - 0.1 * (2,6) = (0.8, 2.4)$

$\Theta^2 = (0.8, 2.4) - 0.1 * (1.6, 4.8) = (0.64, 1.92)$

$\Theta^3 = (0.5124, 1.536)$

$\Theta^4 = (0.4096, 1.22880000000001)$

$\vdots$

$\Theta^{10} = (0.1073741824000003, 0.3221225472000005)$

$\Theta^{50} = \left(1.141798154164342\mathrm{e}^{-05}, 3.42539442494306\mathrm{e}^{-05}\right)$

$\vdots$

$\Theta^{100} = \left(1.6296287810675902\mathrm{e}^{-10}, 4.888886343202771\mathrm{e}^{-10}\right)$

The Gauss-Markov theorem

The baseline multiple regression has the following GMT assumptions:

- In the population,linearity holds so that $y = \beta_0 + \beta_1 x_1 \ldots + \beta_p x_p + \epsilon$
- $x_1, ..., x_p$ satisfies I.I.D, then from the law of large numbers: $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} = [\sum_{i=1}^{N} \boldsymbol{x_i}^\top \boldsymbol{x_i}]^{-1} [\sum_{i=1}^{n} \boldsymbol{x_i}^\top \epsilon_i] \xrightarrow{p} E[\boldsymbol{x_i}^\top \boldsymbol{x_i}]^{-1} E[\boldsymbol{x_i}^\top \epsilon_i] = 0$
- No perfect collinearity : $r[\boldsymbol{X}^\top \boldsymbol{X}] = p$
- $E[\epsilon_i | \boldsymbol{x_i}] = 0 : E[\widehat{\beta} | \boldsymbol{X}] = \beta + \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top E[\boldsymbol{\epsilon} | \boldsymbol{X}] = \beta$
- Homoskedasticity : $E[\epsilon_i^2 | \boldsymbol{x_i}] = \sigma^2 : \widehat{\beta}$ is the **BLUE**

$$\widehat{\beta} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top (\boldsymbol{X}\beta + \boldsymbol{\epsilon}) = \beta + \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon}$$

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_N \end{pmatrix} \qquad \boldsymbol{x}_i = (x_{11}, x_{12}, ..., x_{1p}) \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

$$\boldsymbol{X}^\top = (\boldsymbol{x}_1^\top, ..., \boldsymbol{x}_N^\top)$$

$$\boldsymbol{x}_i^\top \epsilon_i = (x_{i1}\epsilon_i, x_{i2}\epsilon_i, ..., x_{ip}\epsilon_i)$$

How to conduct statistical inference

The key ingredient for statistical inference is a **test statistic with a known distribution** calculated from regression coefficients.
Adding the assumption of **normality**, we have:

$$\widehat{\beta} \sim N(\beta, \left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\sigma^2) \tag{1}$$

$$\widehat{\sigma^2} = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 \tag{2}$$

$$t_j = \frac{\widehat{\beta_j}}{\sqrt{v_j}} \sim t_{N-p-1} \tag{3}$$

$$\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\sigma^2 = \begin{pmatrix} \operatorname{var}(\beta_0) & \operatorname{cov}(\beta_0,\beta_1) & \cdots \\ \operatorname{cov}(\beta_1,\beta_0) & & \vdots \\ \vdots & \ddots & \vdots \\ & \cdots & \operatorname{var}(\beta_{p+1}) \end{pmatrix}$$

How to diagnose and remedy problems

Diagnostics validate the model assumptions and help us prevent wrong conclusions when interpreting the result and conducting statistical inference.They include:

- Goodness of fit
- Remedy for Heteroskedasticity
- Serial correlation
- Multicollinearity

Goodness of fit

Goodness of fit metrics differ in how they measure the fit. For example:

- $\boldsymbol{R}^2 = 1 - \frac{RSS}{TSS}$, and $adjusted\,\boldsymbol{R}^2 = 1 - \frac{RSS/(N-p-1)}{TSS/(N-1)}$
- $\mathrm{AIC} = -2\log(\mathcal{L}^*) + 2p$
  where $\mathcal{L}^*$ is the value of the maximized likelihood function and p is the number of parameters.
- $\mathrm{BIC} = -2\log(\mathcal{L}^*) + log(N)p$

Both metrics penalize for complexity. BIC imposes a higher penalty, so it might underfit relative to vice versa.
In practice, both criteria can be used jointly to guide model selection when the goal is an in-sample fit.

Heteroskedasticity

If the residual variance is correlated with an input variable:

- $\widehat{\beta}$ is unbiased but is not the BLUE
- The statistical inference is not reliable any more.

Remedy for the heteroskedasticity problem under OLS estimates:

- **Robust standard errors** :$\left(X^T X\right)^{-1} X^T \Omega X \left(X^T X\right)^{-1}$
- **Clustered standard errors** : Assume there are distinct groups in the data that are homoscedastic, but the error variance differs between groups

We can also replace OLS by: WLS, GLSAR, GLS.....

Linear Models

The Basis of Multi-Factor Model

The average return of a stock is the payoff for taking risk.

- Average stock return = factor exposure × factor premium:

$$r_i = \alpha_i + \beta_{i1} f_1 + \cdots + \beta_{iK} f_K + \epsilon_i$$

- And the average stock return is the cross-sum of the factor exposure and the factor premium:

$$E(r_i) = E(\alpha_i) + \beta_{i1} E(f_1) + \cdots + \beta_{iK} E(f_K)$$

- For the convenience of exposition, we use vectors to simplify the equation:

$$E(r_i) = \boldsymbol{\beta}_i' E(\mathrm{f})$$

From the CAPM to the Fama-French factor models

- The CAPM model takes it that investors require compensation for systematic risk:

$$E[r_i] - r_f = \alpha_i + \beta_i (E[r_m] - r_f)$$

- The model implies that the value of α_i should be zero, which failed the empirical tests.
- Kenneth French and Eugene Fama identified additional risk factors:
  - **Size**: Market equity (ME)
  - **Value**: Book value of equity (BE) divided by ME
  - **Operating profitability (OP)**: Revenue minus cost of goods sold/assets
  - **Investment**: Investment/assets

How Fama and French Obtain the factors

- The factors result from sorting stocks:

		BM		
		High(30%)	Middle(40%)	Low(30%)
ME	Small(50%)	S/H	S/M	S/L
	Big(50%)	B/H	B/M	B/L

- Then we get two factors:

$$SMB_t = \frac{(SL_t + SM_t + SH_t)}{3} - \frac{(BL_t + BM_t + BH_t)}{3}$$

$$HML_t = \frac{(BH_t + SH_t)}{2} - \frac{(BL_t + SL_t)}{2}$$

- Finally we run regression for risk exposure β_{HML} and β_{SMB}

Time Series Regression

- If risk factors are returns of a certain portfolio itself, we can use time series regression.
- Take CAPM model as an example:

ticker	time	$R_A - R_f$	$R_M - R_f$
A	2018	0.02	0.013
A	2019	-0.03	0.002
A	2020	0.02	0.02
A	2021	0.015	0.021

- We can run time series regressions to get the factor exposure:

$$R_{it} = \alpha_i + \beta_i' f_t + \varepsilon_{it}$$
$$\mathrm{E}_T[R_i] = \alpha_i + \beta_i' \mathrm{E}_T[f_t]$$

Cross-Section Regression

ticker	time	return	GDP(trillion)
A	2019	0.005	14
A	2020	0.003	14
A	2021	-0.002	17
B	2019	0.002	14
B	2020	-0.001	14
B	2021	-0.003	17
C	2019	0.012	14
C	2020	-0.002	14
C	2021	0.012	17

If risk factors are not returns themselves, we need one more step to get risk premium:

$$R_{it} = a_i + \beta_i' \lambda_t + \varepsilon_{it}$$
$$\mathrm{E}_T[R_i] = \alpha_i + \beta_i' \mathrm{E}_T[f_t]$$

Fama-Macbeth regression

ticker	time	return	GDP (trillion)	β
A	2019	0.005	14	β_A
A	2020	0.003	14	β_A
A	2021	-0.002	17	β_A
B	2019	0.002	14	β_B
B	2020	-0.001	14	β_B
B	2021	-0.003	17	β_B
C	2019	0.012	14	β_C
C	2020	-0.002	14	β_C
C	2021	0.012	17	β_C

First stage:

$$R_{it} = a_i + \beta_i' f_t + \varepsilon_{it}$$

Second stage:

$$R_{it} = \beta_i' \lambda_t + \alpha_{it}$$

How FM Regression Address the Inference Problem

If we use pool regression:

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}y_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^2} \\ &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\left(x_{it}\beta + \varepsilon_{it}\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^2} \\ &= \beta + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\varepsilon_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^2}\end{aligned}$$

How FM Regression Address the Inference Problem

$$
\begin{aligned}
\operatorname{Var}\left[\beta_{ols}\right] &= E\left(\left[\sum_{t=1}^{T}\sum_{i=1}^{N} x_{it}\varepsilon_{it}\right]^2\left[\sum_{t=1}^{T}\sum_{i=1}^{N} x_{it}^2\right]^{-2}\right) \\
&= E\left(\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{N} x_{it}\varepsilon_{it}\right)^2\right]\left[\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^2\right]^{-2}\right) \\
&= E\left(\left[\sum_{t=1}^{T}\left(\sum_{i=1}^{N} x_{it}^2\varepsilon_{it}^2 + 2\sum_{i=1}^{N}\sum_{s=i+1}^{N} x_{it}x_{st}\varepsilon_{it}\varepsilon_{st}\right)\right]\left[\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^2\right]\right] \\
&= \left(NT\sigma_x^2\sigma_\varepsilon^2 + TN(N-1)\rho_x\sigma_x^2\rho_\varepsilon\sigma_\varepsilon^2\right)/\left(NT\sigma_x^2\right)^2 \\
&= \frac{\sigma_\varepsilon^2}{NT\sigma_x^2}\left(1+(N-1)\rho_x\rho_\varepsilon\right)
\end{aligned}
$$

How FM Regression Address the Inference Problem

If we use cross-section regression:

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{N} x_i E_T(y_i)}{\sum_{i=1}^{N} x_i^2} \\ &= \frac{\sum_{i=1}^{N} x_i (x_i \beta + \varepsilon_i)}{\sum_{i=1}^{N} x_i^2} \\ &= \beta + \frac{\sum_{i=1}^{N} x_i \varepsilon_i}{\sum_{i=1}^{N} x_i^2}\end{aligned}$$

How FM Regression Address the Inference Problem

$$
\begin{aligned}
\text{Var}\left[\beta_{ols}\right] &= E\left(\hat{\beta}_{ols} - \beta\right)^2 \\
&= E\left(\left[\sum_{i=1}^{N} x_i \varepsilon_i\right]^2 \left[\sum_{i=1}^{N} x_i^2\right]^{-2}\right) \\
&= E\left(\left[\sum_{i=1}^{N} x_i^2 \varepsilon_i^2 + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} x_i x_j \epsilon_i \epsilon_j\right] \left[\sum_{i=1}^{N} x_i^2\right]^{-2}\right) \\
&= \left(N\sigma_x^2 \sigma_\varepsilon^2 + N(N-1)\rho_x \sigma_x^2 \rho_\epsilon \sigma_\epsilon^2\right) / \left(N\sigma_x^2\right)^2 \\
&= \frac{\sigma_\varepsilon^2}{N\sigma_x^2}\left(1 + (N-1)\rho_x \rho_\epsilon\right)
\end{aligned}
$$

How FM Regression Address the Inference Problem

If we use Fama-Macbeth regression:

$$\begin{aligned}\hat{\beta}_{FM} &= \sum_{t=1}^{T} \frac{\hat{\beta}_t}{T} \\ &= \frac{1}{T} \sum_{t=1}^{T} \left(\frac{\sum_{i=1}^{N} x_{it} y_{it}}{\sum_{a}^{b} x_{it}^2} \right) \\ &= \beta + \frac{1}{T} \sum_{t=1}^{T} \left(\frac{\sum_{i=1}^{N} x_{it} \varepsilon_{it}}{\sum_{a}^{b} x_{it}^2} \right)\end{aligned}$$

How FM Regression Address the Inference Problem

$$
\begin{aligned}
\operatorname{Var}\left(\hat{\beta}_{FM}\right) &= \frac{1}{T^2}\operatorname{Var}\left(\sum_{t=1}^{T}\hat{\beta}_t\right) \\
&= \frac{\operatorname{Var}\left(\hat{\beta}_t\right)}{T} + \frac{2\sum_{t=1}^{T-1}\sum_{s=t}^{T}\operatorname{Cov}\left(\hat{\beta}_t,\hat{\beta}_s\right)}{T^2} \\
&= \frac{\operatorname{Var}\left(\hat{\beta}_t\right)}{T} + \frac{T(T-1)}{T^2}\operatorname{Cov}\left(\hat{\beta}_t,\hat{\beta}_s\right) \\
&= \frac{1}{T}\left(\frac{\sigma_\epsilon^2}{N\sigma_x^2}\right) + \frac{T(T-1)}{T^2} * 0 \\
&= \frac{\sigma_\epsilon^2}{NT\sigma_x^2}
\end{aligned}
$$

How FM Regression Address the Inference Problem

- The stand error from pool regression is:

$$\text{Var}\left[\beta_{ols}\right] = \frac{\sigma_\varepsilon^2}{N\sigma_x^2}\left(1 + (N-1)\rho_x\rho_\epsilon\right)$$

- The stand error from cross-section is:

$$\text{Var}\left[\beta_{ols}\right] = \frac{\sigma_\varepsilon^2}{N\sigma_x^2}\left(1 + (N-1)\rho_x\rho_\epsilon\right)$$

- The stand error from Fama-Macbeth is:

$$\text{Var}\left(\hat{\beta}_{FM}\right) = \frac{\sigma_\epsilon^2}{NT\sigma_x^2}$$

Barra Models

ticker	time	return	P/E
A	2019	0.005	23.27
A	2020	0.003	25.84
A	2021	-0.002	27.6
B	2019	0.002	58.89
B	2020	-0.001	40
B	2021	-0.003	49.77
C	2019	0.012	15.47
C	2020	-0.002	13.76
C	2021	0.012	16.9

In the Barra model, some of the factor exposure is the value of some observable (or easily calculable) characteristic of the stock.

$$r_i = \alpha_i + \beta_{i1} f_1 + \cdots + \beta_{iK} f_K + \epsilon_i$$

Linear Models

Regularizing linear regression using shrinkage

When a linear regression model contains many correlated variables, their coefficients will be poorly determined, and the model will more likely to overfit the sample.

- In OLS, β is estimated like this:

$$\widehat{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T y$$

- Notice that $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ equals $\frac{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^*}{\left|\boldsymbol{X}^T\boldsymbol{X}\right|}$

Shrinkage models

One popular technique to control overfitting is regularization, which involves the addition of a penalty term to the error function to discourage the coefficients from reaching large values.

$$\widehat{\beta}^{S} = \underset{\beta^{S}}{\operatorname{argmin}} \sum_{i=1}^{N} \left[\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 + \lambda S(\beta) \right]$$

Ridge Regression

- The penalty term:

$$S(\beta) = \sum_{i=1}^{p} \beta_i^2 = \|\beta\|^2$$

- Hence, the ridge coefficients are defined as:

$$\begin{aligned}\widehat{\beta}^{\text{Ridge}} &= \underset{\beta^{\text{Rdge}}}{\text{argmin}} \sum_{i=1}^{N} \left[\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right] \\ &= \underset{\beta^{\text{Ridge}}}{\text{argmin}} (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda \beta^T \beta\end{aligned}$$

Ridge Regression

- The intercept β_0 as been excluded from the penalty to make the procedure independent of the origin chosen for the output variable.
- It is important to standardize the inputs. This is because the ridge solution is sensitive to the scale of the inputs.
- There is also a closed solution for the ridge estimator that resembles the OLS case:

$$\widehat{\boldsymbol{\beta}}^{\text{Ridge}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Lasso Regression

The lasso penalty is the sum of the absolute values of the coefficient vector.

- Hence, the lasso coefficients are defined as:

$$\begin{aligned}\widehat{\beta}^{\text{Lasso}} &= \underset{\beta^{\text{Lasso}}}{\text{argmin}} \sum_{i=1}^{N} \left[\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] \\ &= \underset{\beta^{\text{Lasso}}}{\text{argmin}} (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda \|\beta\|_1\end{aligned}$$

Lasso Regression

- Similar to ridge regression, the inputs need to be standardized. The lasso penalty makes the solution nonlinear.
- The lasso penalty had the effect of gradually reducing some coefficients to **zero** as the regularization increases. For this reason, the lasso can be used for the continuous selection of a subset of features.

Logistic Regression

- The output variable y:

$$y_t = \begin{cases} 1 & A \\ 0 & \text{otherwise} \end{cases}$$

- Logistic regression models the probability that y belongs to either of the categories.

$$P(\boldsymbol{x}_t) = P_r(y_t = 1 \mid \boldsymbol{x}_t).$$

Logistic Regression

To prevent the model from producing values outside the [0, 1] interval, we must model p(x) using a function that only gives outputs between 0 and 1 over the entire domain of x.

$$p(x) = \frac{1}{1 + e^{-x\beta}}$$

Logistic Regression

More formally, we are seeking to maximize the likelihood function:

$$\max_{\beta} \mathcal{L}(\beta) = \prod_{i:y_i=1} p(\boldsymbol{x}_i) \prod_{i':y_{i'}=0} (1 - p(\boldsymbol{x}_{i'}))$$

$$\beta^{\mathrm{ML}} = \operatorname{argmax} \log \mathcal{L}(\beta)$$

$$i.e \sum_{i=1}^{N} (y_i \log p(\boldsymbol{x}_i, \beta) + (1 - y_i) \log (1 - p(\boldsymbol{x}_i, \beta)))$$

Confusion matrix

		Actual (Truth)	
		Positive	**Negative**
Prediction	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$\text{Accuracy} = \frac{\#\text{ Correct Predictions}}{\#\text{ Cases}} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{True Positive Rate (Sensitivity, Recall)} = \frac{\#\text{ Correct Positive Predictions}}{\#\text{ Positive Cases}} = \frac{TP}{TP + FN}$$

$$\text{False Negative Rate (Miss Rate)} = 1 - \text{True Positive Rate}$$

$$\text{True Negative Rate (Specificity)} = \frac{\#\text{ Correct Negative Predictions}}{\#\text{ Negative Cases}} = \frac{TN}{TN + FP}$$

$$\text{False Positive Rate (Fall-Out)} = 1 - \text{True Negative Rate}$$

Figure: Confusion matrix and related error metrics

ROC and AUC

- Receiver operating characteristics (ROC) curve allows us to visualize and select classifiers based on their performance.
- The area under the curve (AUC) is defined as the area under the ROC plot that varies between 0.5 and the maximum of 1.

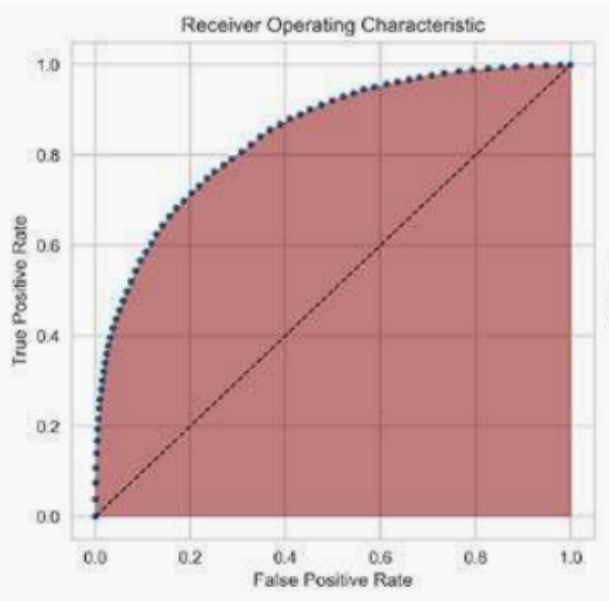


Figure: Receiver operating characteristics (ROC) curve