Machine Learning for Finance (FIN 570)
Regression

Instructor: Jaehyuk Choi

Peking University HSBC Business School, Shenzhen, China

2018-19 Module 1 (Fall 2018)

Notations and conventions: vector and matrix

General rules (guess from the context)

- Scalar (non-bold): x, y, X, Y
- Vector (lowercase bold): $\boldsymbol{x} = (x_i), \boldsymbol{y} = (y_i)$
- Matrix (uppercase bold): $\boldsymbol{X} = (X_{ij}), \boldsymbol{Y} = (Y_{ij})$
- The (i, j) component of $\boldsymbol{X}$: X_{ij}
- The i-th row vector of $\boldsymbol{X}$: $\boldsymbol{X}_{i*} = (X_{i1}, X_{i2}, \cdots, X_{ip})^T$
- The j-th column vector of $\boldsymbol{X}$: $\boldsymbol{X}_{*j} = (X_{1j}, X_{2j}, \cdots, X_{Nj})$

Examples

- Dot product: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y}$
- Vector norm: $|\boldsymbol{x}| = \sqrt{\boldsymbol{x}^T \boldsymbol{x}}$
- Matrix multiplication: $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{Y} \rightarrow Z_{ij} = \boldsymbol{X}_{i*}\boldsymbol{Y}_{*j}$

Notation and conventions: variables and observations

General rules

- Generic (or representative) variables (uppercase non-bold): X (input), Y (output), G (classification output)
- The predictions: $\hat{Y}$, $\hat{G}$
- X (input) may be p-dimensional (features/predictors): X_j $(j \le p)$, row vector
- Y (output) may be K-dimensional (responses): Y_k $(k \le K)$, row vector.
- The N observations of X or Y are stacked over as rows: $\boldsymbol{X}$ $(N \times p$ matrix), $\boldsymbol{Y}$ $(N \times K$ matrix)
- The i-th observation of the j-th feature: X_{ij} v.s. $X_j^{(i)}$ in **PML**
- The i-th observation set: $\boldsymbol{X}_{i*}$ $(1 \times p$ row vector) v.s. $X^{(i)}$ in **PML**
- All observation of j-th feature X_j: $\boldsymbol{X}_{*j}$ $(N \times 1$ row vector) v.s. X_j in **PML**
- $\boldsymbol{X} = (\boldsymbol{X}_{*1} \cdots \boldsymbol{X}_{*p})$ (column-wise concatenation)
- The weight vector, $\boldsymbol{\beta}$ or $\boldsymbol{w}$, are column vectors used interchangeably.

Simple Linear Regression (Ordinary Least Square)

For scalar predictor (X) and response (Y),

$$Y \approx \beta_0 + \beta_1 X \quad \longrightarrow \quad \hat{\boldsymbol{y}} = \beta_0 + \beta_1 \boldsymbol{x}.$$

For N observations $(x_1, y_1), \cdots, (x_N, y_N)$, the set of $(\hat{\beta}_0, \hat{\beta}_1)$ to minimize the residual sum of squares (RSS):

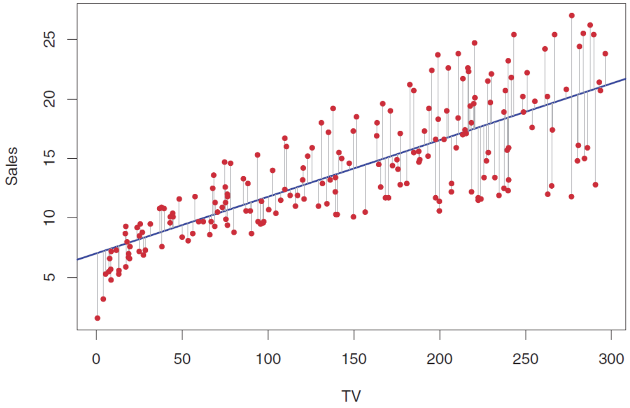
$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 = (\boldsymbol{y} - \beta_0 - \beta_1 \boldsymbol{x})^T (\boldsymbol{y} - \beta_0 - \beta_1 \boldsymbol{x})$$

is given as

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

for $\bar{x} = \sum x_i / N$ and $\bar{y} = \sum y_i / N$.

Figure 3.1 (p62) from **ISLR**



Multi-dimensional Linear Regression

For $(p+1)$-vector predictor (X) and scalar response (Y),

$$Y \approx X\boldsymbol{\beta} \quad \longrightarrow \quad \hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta},$$

where $X_0 = 1$ ($\boldsymbol{X}_{*0} = \mathbf{1}$) and $\boldsymbol{\beta}$ is a $(p+1)$-column vector.

$$\mathsf{RSS}(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\mathsf{RSS}(\boldsymbol{\beta}) = -\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

For $(p+1)$-vector predictor (X) and K-vector response (Y), the result is similarly given as

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{B} \quad \text{where} \quad \hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y},$$

which is the independent regressions on Y_j ($\boldsymbol{Y}_{*j}$) combined together,

$$\hat{\boldsymbol{B}}_{*j} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_{*j}$$

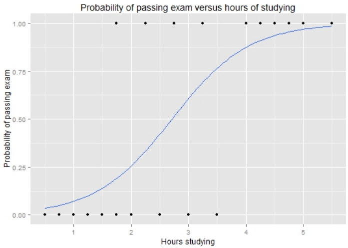
Logistic Regression (Classification)

- Qualitative (categorical) response (binary dependent variable, $Y \in \{0,1\}$)
- Multiple categories: how to give order?
- Linear regression (quantitative) is not proper
- Logistic (sigmoid) function: $\sigma(\text{logit}) = \text{quantile}$

$$p = \phi(t) = \frac{e^t}{1+e^t} = \frac{1}{1+e^{-t}} \quad \text{for} \quad t = Xw \; (X_0 = 1)$$

- Logit function (the inverse): log odds

$$\phi^{-1}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p)$$



Fitting of logistic regression

Likelihood function

- For a given the prediction model, measures the likelihood of a data set.
- The best prediction model/weight is the one that maximizes the likelihood of the dataset.

For a data set $(\boldsymbol{X}, \boldsymbol{y})$ where $y_i \in \{0, 1\}$),

$$L(\boldsymbol{w}) = \prod_i P(y_i = \hat{y}_i) = \prod_{i:y_i=1} \phi(\boldsymbol{X}_{i*}\boldsymbol{w}) \prod_{i:y_i=0} \big(1 - \phi(\boldsymbol{X}_{i*}\boldsymbol{w})\big)$$

$$= \prod_i \phi(\boldsymbol{X}_{i*}\boldsymbol{w})^{y_i} \big(1 - \phi(\boldsymbol{X}_{i*}\boldsymbol{w})\big)^{1-y_i}$$

$$\log L(\boldsymbol{w}) = \sum_i y_i \log \phi(\boldsymbol{X}_{i*}\boldsymbol{w}) + (1 - y_i) \log \big(1 - \phi(\boldsymbol{X}_{i*}\boldsymbol{w})\big)$$

The cost function (to minimize) is $J(\boldsymbol{w}) = -\log L(\boldsymbol{w})$