Linear Regression

1. Linear Regression Model

1.1 Model & Notations

$x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T \in \mathbb{R}^p$

$y = (y_1, y_2, \cdots, y_N)^T$, $X = (x_1, x_2, \cdots, x_N)^T \in \mathbb{R}^{N\times p}$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N)^T \in \mathbb{R}^N$

$y = X\beta + \varepsilon$

remarks:

1. Quantitative inputs & its transformations (log, squares) & basis expansions ($X_2 = X_1^2$, $X_3 = X_1^3$)
2. Qualitative inputs: dummy variable coding
3. Interaction between variables ($X_3 = X_1 \cdot X_2$)

1.2 Model Assumptions

(A1) The relationship between response y and covariates X is linear

(A2) X is non-stochastic matrix and $\text{rank}(X) = p$

(A3) $E(\varepsilon) = 0$. This implies $E(y) = X\beta$

(A4) $\text{cov}(\varepsilon) = E(\varepsilon\varepsilon^T) = \sigma^2 I_N$

(A5) ε follows multivariate normal distribution $N(0, \sigma^2 I_N)$

or

(A_2^*) X is a full rank matrix with probability 1 ($\lambda_{min}(X^TX) \to \infty$ a.s.)

(A_3^*) $E(\varepsilon | X) = 0$

(A_4^*) $E(\varepsilon\varepsilon^T | X) = \sigma^2 I_N$

(A_5^*) $\varepsilon | X \sim N(0, \sigma^2 I_N)$

2. Model Estimation

· OLS estimation: $RSS(\beta)=\sum_{i=1}^{N}\{y_i-f(x_i)\}^2=\sum_{i=1}^{N}\{y_i-\beta_0-\sum_j x_{ij}\beta_j\}^2$

$$=(y-X\beta)^T(y-X\beta)$$

This criterion is valid if y_i's are conditionally independent given inputs x_i

$$\frac{\partial RSS(\beta)}{\partial\beta}=-2(y-X\beta)^TX=0,\quad \hat\beta=(X^TX)^{-1}X^Ty,\quad \hat Y=X(X^TX)^{-1}X^Ty=Hy$$

1. Assume X is full rank, hence X^TX is positive definite
2. Fitted Values: $\hat y=Hy$. Residual vector $y-\hat y$ is orthogonal to the column space of X
3. The residual sum of squares $RSS(\beta)$ can be used as a goodness-of-fit measure.

3. Statistical Inference

3.1 Mean and Variance of the OLS Estimator

$E(\hat\beta)=\beta,\ cov(\hat\beta)=(X^TX)^{-1}\sigma^2$

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

<u>Theorem 1</u>: (Gauss Markov Theorem) Assume (A1)~(A4). Then $\hat\beta$ is the best linear unbiased estimator (BLUE), provided it exists.

It implies, $\hat\beta$ has the smallest variance over all linear unbiased estimator $\tilde\beta$, i.e. $\tilde\beta=\sum_{i=1}^{N}w_iy_i$ and $E(\tilde\beta)=\beta$, $\forall\eta\in R^p$, $\|\eta\|=1$, $Var(\eta^T\hat\beta)\le Var(\eta^T\tilde\beta)$

proof: $\eta^T\hat\beta=\underbrace{\eta^T(X^TX)^{-1}X^T}_{\alpha'}y$ $\quad E(\eta^T\hat\beta)=\eta^T(X^TX)^{-1}X^TE(y)=\eta^T\beta$ $\quad Cov(\eta^T\hat\beta)=\alpha^T\alpha\cdot\sigma^2$

Alternative estimator: $d'Y$, $E(d'Y)=(X^Td)'\beta$ $\quad\therefore X^Td=\eta$

$Cov(d^TY)=d'd\cdot\sigma^2=\sigma^2(d-a+a)'(d-a+a)$ $\quad a'(d-a)=\eta^T(X^TX)^{-1}X^Td-\eta^T(X^TX)^{-1}\eta$

$=\sigma^2(a'a+(d-a)'(d-a)+2a'(d-a))$ $\quad =\eta^T(X^TX)^{-1}\eta-\eta^T(X^TX)^{-1}\eta$

$\geq\sigma^2a'a$ $\quad =0$

property 1 : $(N-p)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p}$

proof $RSS = (y-\hat{y})'(y-\hat{y}) = y'(I-H)y$

$tr(I-H) = tr(I) - tr(X(X'X)^{-1}X') = N-p$

$\exists U$, s.t. $U'(I-H)U = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 0 & \\ & & & 0 \end{pmatrix}$

$RSS = (U'y)' D (U'y) \sim \chi^2_{N-p}$

property 2: $\sum_{j=1}^{n}(y_j-\bar{y})^2 = \sum_{j=1}^{n}(\hat{y}-\bar{y})^2 + \sum_{j=1}^{n}(y_j-\hat{y})^2$

proof $\sum_{j=1}^{n}(y_j-\bar{y})^2 = \sum_{j=1}^{n}(y_j-\hat{y}+\hat{y}-\bar{y})^2$

$$= \sum_{j=1}^{n}(y_j-\hat{y}_j)^2 + \sum_{j=1}^{n}(\hat{y}_j-\bar{y})^2 + 2\sum_{j=1}^{n}(\hat{y}_j-\bar{y})(y_j-\hat{y}_j)$$

$$\sum_{j=1}^{N}(\hat{y}_j-\bar{y})^T(\hat{y}_j-y_j) = y^T(H-\tfrac{1}{n}11^T)(I-H)y$$

$$= y^T(\tfrac{1}{n}11^T - \tfrac{1}{n}11^T)y = 0$$

3.2 Sampling property

- $\hat{\beta} \sim N(\beta, (X^TX)^{-1}\sigma^2)$, $(N-p)\hat{\sigma}^2 \sim \sigma^2\chi^2_{N-p}$
- $z_j = \dfrac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} \sim t_{N-p}$
- $F = \dfrac{(RSS_0 - RSS_1)/(p_1-p_0)}{RSS_1/(N-p)} \sim F(p_1-p_0, N-p_1)$

4. Goodness-of-fit

- $R^2 = 1 - \dfrac{RSS}{TSS}$, Adjusted $R^2 = 1 - \dfrac{RSS/(n-p)}{TSS/(n-1)}$

5. Model Selection

1. Subset Selection

1.1 Best-subset Selection : time-consuming

1.2 Forward-stepwise selection (greedy algorithm): Starts with the intercept, and then sequentially adds into the model the predictor that most improves the fit

1.3 Backward-stepwise selection: starts with the full model, and sequentially deletes the predictor that has the least impact on the fit. $(N>p)$

1.4 Stepwise selection: Consider both forward and backward moves at each step, and select the "best"

$$AIC = -\frac{1}{N}\mathcal{L}(\beta) + 2\frac{d}{N}$$

$$BIC = -2\mathcal{L}(\beta) + (\log N)d$$

Best model has smallest AIC

comment:

1. BIC can consistently select the true model
2. other criterion including C_p

2. Shrinkage Methods

2.1 Ridge Regression

- $$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_j x_{ij}\beta_j\right)^2 + \lambda\sum_j \beta_j^2\right\}$$
$$= (X^TX + \lambda I)^{-1}X^Ty$$
- $$RSS(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

2.2 Lasso Regression

- $$\hat{\beta}^{Lasso} = \underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_j x_{ij}\beta_j\right)^2 + \lambda\sum_j |\beta_j|\right\}$$