Linear Regression Model

- form: $f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$

X_j Comes from:
· input
· transformation of input (log, square-root, square etc)
· basis expasion $X_i^2, X_i^3 \Rightarrow$ polynomial
· "dummy" Coding $(0,1,0,0,0)^T$
· interaction between variables, $X_3 = X_1 \cdot X_2$

Properties

- $RSS(\beta) = \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$

$$= (y - X\beta)'(y - X\beta)$$

$$\frac{\partial RSS}{\partial \beta} = -2X^T(y - X\beta) \qquad \frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = 2X^T X \qquad \hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{\text{Hat Matrix}} y$$

- $Var(\hat{\beta}) = (X^T X)^{-1} \sigma^2$
- $\hat{\sigma}^2 = \frac{1}{N-p-1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$, $(N-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p-1}$
- $Y = E(Y \mid X_1, \cdots, X_p) + \varepsilon = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$

Hypothesis Testing

- t-test

$$H_0: \beta_j = 0, \quad H_1: \beta_j \neq 0$$

Z-score: $z_j = \dfrac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$ $\quad v_j$: jth diagonal element of $(X^TX)^{-1}$

$z_j \sim t_{N-p-1}$ or $N(0,1)$ (if σ is known) $\quad \hat{\sigma}^2 = RSS/(n-p)$

- F-test

$H_0: \beta_1 = \beta_2 = \cdots = \beta_q = 0 \quad H_1$: at least one of $\beta_1 \sim \beta_q \neq 0$

F-stat: $F = \dfrac{(RSS_0 - RSS_1)/(p-q)}{RSS_1/(N-p-1)}$ $\quad RSS_1$: LS fit of $p+1$ variables, RSS_0: LS fit of $q+1$ variables

$F \sim F_{p-q,n-p-1}$ or $\chi^2_{p_1-p_0}$ (for large N)

Q: Why $F \sim F_{p-q,n-p-1}$? What is exercise 3.1?

- Confidence interval:

$\beta_j \in (\hat{\beta}_j - z^{(1-\alpha)} v_j^{\frac{1}{2}}\hat{\sigma}, \hat{\beta}_j + z^{(1-\alpha)} v_j^{\frac{1}{2}}\hat{\sigma})$ for single β_j

$C_\beta \in \{\beta \mid (\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta) \leq \hat{\sigma}^2 \chi^{2\,(1-\alpha)}_{p+1}\}$ for entire β

PROOF: Let $V = (X^TX)^{1/2}(\hat{\beta}-\beta)$, then

$$\begin{cases} E(V) = 0 \\ Cov(V) = (X'X)^{1/2} Cov(\hat{\beta})(X'X)^{1/2} = \sigma^2 I \end{cases}$$

since V is normally-distributed, $V'V \sim \sigma^2 \chi^2_{r+1}$

$\Rightarrow (\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) \sim \sigma^2\chi^2_{r+1}$, and $(n-r-1)s^2 = \hat{\varepsilon}'\hat{\varepsilon} \sim \sigma^2\chi^2_{n-r-1}$

$\Rightarrow [V'V/(r+1)]/s^2$ is $F_{r+1,n-r-1}$ distribution

$\Rightarrow (\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) \leq (r+1)\cdot s^2 F_{r+1,n-r-1}(\alpha)$

My idea: $(X'X)^{-\frac{1}{2}}(\beta-\hat{\beta}) \perp\!\!\!\perp \hat{\varepsilon}$ but why?

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

The Gauss-Markov Theorem

- **Gauss-Markov Theorem:** $c'\hat{\beta} = c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \cdots + c_r\hat{\beta}_r$ is the smallest possible variance among all linear estimators of $c'\beta = c_0\beta_0 + \cdots + c_r\beta_r$

PROOF $c'\beta = c_0\beta_0 + \cdots + c_r\beta_r$

Estimator: $c'\hat{\beta} = c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \cdots + c_r\hat{\beta}_r = \underbrace{c'(Z'Z)^{-1}Z'}_{a'}Y$

$E(c'\hat{\beta}) = c'\beta$, $V(c'\hat{\beta}) = a'\sigma^2 I a = \sigma^2 a'a$

Alternative Estimator: $d'Y$ with $E(d'Y) = d'Z\beta = c'\beta$, $\forall\beta \Leftrightarrow d'Z = c'$

$$V(d'Y) = d'\sigma^2 I d = \sigma^2 d'd = \sigma^2[(a+d-a)'(a+d-a)]$$

$$= \sigma^2[a'a + (d-a)(d-a) + 2a'(d-a)] = \sigma^2[a'a + (d-a)'(d-a)]$$

$$\geq V(c'\hat{\beta})$$

$$a'(d-a) = c'(Z'Z)^{-1}Z'(Z'^{-1} - Z(Z'Z)^{-1})c$$

$$= c'(Z'Z)^{-1}c - c'(Z'Z)^{-1}c = 0$$

$\Rightarrow$ $c'\hat{\beta}$ is the BLUE of $c'\beta$

- **Mean Square Error:** $MSE(\tilde{\theta}) = E(\tilde{\theta}-\theta)^2 = Var(\tilde{\theta}) + [E(\tilde{\theta})-\theta]^2$

opt. ↑ ↑ zero

B.L.U.E.

Multiple Outputs

- $Y_K = \beta_{0K} + \sum_{j=1}^{p} X_j \beta_{jK} + \varepsilon_K = f_K(X) + \varepsilon_K$

  $Y = XB + E$

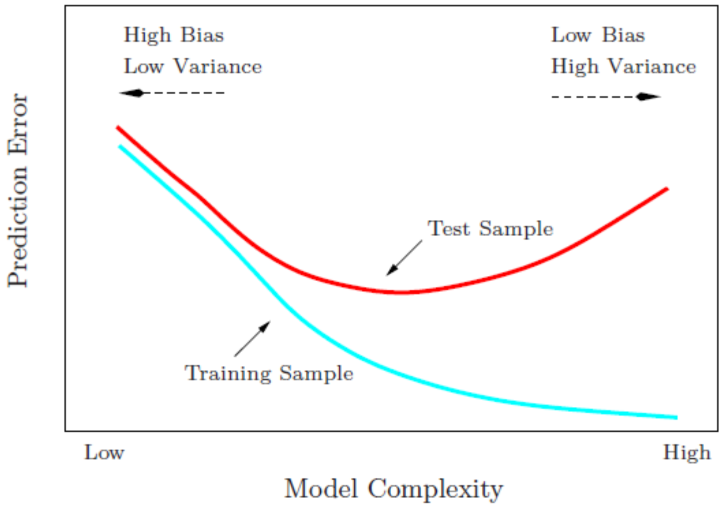
- $RSS(B) = \sum_{K=1}^{K} \sum_{i=1}^{N} (y_{iK} - f_K(x_i))^2 = tr[(Y-XB)^T(Y-XB)]$

  $\hat{B} = (X^T X)^{-1} X^T Y$

- if $\varepsilon_1, \cdots, \varepsilon_K$ are correlated, $Cov(\varepsilon) = \Sigma$

  $RSS(B; \Sigma) = \sum_{i=1}^{N} (y_i - f(x_i))^T \Sigma^{-1} (y_i - f(x_i))$

Subset Selection discrete



- Forward-stepwise v.s. Backward-stepwise

Forward-stepwise → start with β_0 → add params → improve the fit

Backward-stepwise → drop the smallest Z-score → reduce the variance

- AIC/BIC Criterion: ??

$$AIC = -\frac{2}{N} l(\beta) + 2\frac{d}{N}, \quad BIC = -2l(\beta) + (\log N)d$$

· Shrinkage Methods continuous

· Ridge Regression

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

or

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \le t$$

?Ex. 3.5 Consider the ridge regression problem (3.41). Show that this problem is equivalent to the problem

$$\hat{\beta}^c = \underset{\beta^c}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left[y_i - \beta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j^c \right]^2 + \lambda \sum_{j=1}^{p} \beta_j^{c2} \right\}. \tag{3.85}$$

Give the correspondence between β^c and the original β in (3.41). Characterize the solution to this modified criterion. Show that a similar result holds for the lasso.

· $RSS(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$

$$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1} X^T y$$

· MLE v.s. MAP

$$\hat{\theta}_{MLE} = \operatorname{argmax} P(X;\theta) = \operatorname{argmax} \sum_{i=1}^{n} \log P(x_i;\theta) = \operatorname{argmin} -\sum_{i=1}^{n} \log P(x_i;\theta)$$

$$\begin{aligned}\hat{\theta}_{MAP} &= \operatorname{argmax} P(\theta|X) \\ &= \operatorname{argmin} -\log P(\theta|X) \\ &= \operatorname{argmin} -\log(X|\theta) - \log P(\theta) + \underbrace{\log P(X)}_{=const} \\ &= \operatorname{argmin} -\log P(X|\theta) + \tfrac{1}{2}\theta'\Sigma^{-1}\theta\end{aligned}$$

since $\theta \sim N(0, \Sigma)$,

$$P(\theta) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}\theta'\Sigma^{-1}\theta}$$

· log-posterior of β: $\tilde{l}(\beta) = -\sum \log P(y_i;\beta) + \log P(\beta)$

$$= -\sum_{i=1}^{n} (y_i - \beta_0 - x_i^T\beta)^2 / 2\sigma^2 + \frac{1}{2\tau^2}\sum_{i=1}^{p}\beta_i^2$$

$$\Leftrightarrow -\sum_{i=1}^{n} (y_i - \beta_0 - x_i^T\beta) + \underbrace{\frac{\sigma^2}{\tau^2}}_{=\lambda} \sum_{i=1}^{p} \beta_i^2$$