LINEAR MODEL SELECTION

23/03/21

Standard linear model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

Many generalisations: glm, polynomial regression, spline models (chapter 7)

Standard models are used very often as they are easily interpretable and have good predictive accuracy.

β is estimated via least-squares

LEAST - SQUARES

Data : $(y_i, x_{i1}, x_{i2}, \dots, x_{ip})$, $i = 1, \dots, n$

In matrix form :

$$\underline{Y} = X \underline{\beta} + \underline{\varepsilon} \quad , \quad \underline{\varepsilon} \sim N(0, \sigma^2 I_n)$$

$\underline{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$

$\underline{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $\quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & & x_{np} \end{pmatrix}$ $\quad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

$\begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix}$

LS $\quad \min_{\underline{\beta}} \sum_{i=1}^{n} (y_i - x_i^t \underline{\beta})^2$ $\quad$ where $\underline{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$

$= \min_{\underline{\beta}} \underbrace{(\underline{Y} - X\underline{\beta})^t}_{1 \times n} \underbrace{(\underline{Y} - X\underline{\beta})}_{n \times 1}$

$(a+b)(c+d)$

$$\min_{\beta} (\underline{y} - X\underline{\beta})^t (\underline{y} - X\underline{\beta})$$

$$= \underline{y}^t\underline{y} - \underline{y}^t X\underline{\beta} - (X\underline{\beta})^t \underline{y} + (X\underline{\beta})^t (X\underline{\beta})$$

$$(AB)^t = B^t A^t$$

$$= \underline{y}^t\underline{y} - \underline{y}^t X\underline{\beta} - \underline{\beta}^t X^t \underline{y} + \underline{\beta}^t X^t X\underline{\beta}$$

$$= \underline{y}^t\underline{y} - 2\underline{\beta}^t \underset{p\times n}{X^t} \underset{n\times 1}{\underline{y}} + \underline{\beta}^t X^t X\underline{\beta}$$

$$\frac{d}{d\underline{\beta}} (\underline{y} - X\underline{\beta})^t (\underline{y} - X\underline{\beta}) = -2X^t\underline{y} + 2X^tX\underline{\beta} = 0$$

p+1 equations in p+1 parameters (NORMAL EQUATIONS)

$$(X^tX)\underline{\beta} = X^t\underline{y}$$

$$\boxed{\hat{\underline{\beta}} = (X^tX)^{-1}X^t\underline{y}}$$

$\hat{\beta}$ PROPERTIES

① $$E[\hat{\underline{\beta}}] = E[(X^tX)^{-1}X^t\underline{y}] = (X^tX)^{-1}X^t E[\underline{y}]$$

$$= \underbrace{(X^tX)^{-1}X^tX}_{I}\underline{\beta} = \underline{\beta}$$

UNBIASED

② $\mathrm{Var}(\hat{\underline{\beta}})$ = covariance matrix

$$= \mathrm{Var}((X^tX)^{-1}X^t\underline{y}) = (X^tX)^{-1}X^t\,\mathrm{Var}(\underline{y})\,X(X^tX)^{-1}$$

$\mathrm{Var}(A\underline{y}) = A\,\mathrm{Var}(\underline{y})A^t$

$$= (X^tX)^{-1}X^t\sigma^2 I X(X^tX)^{-1}$$

$$= \sigma^2\underbrace{(X^tX)^{-1}X^tX}_{I}(X^tX)^{-1} = \sigma^2(X^tX)^{-1}$$

COVARIANCE MATRIX

$$\begin{pmatrix} \mathrm{var}(\hat{\beta}_0) & \mathrm{cov}(\hat{\beta}_0,\hat{\beta}_1) & \cdots & \mathrm{cov}(\hat{\beta}_0,\hat{\beta}_p) \\ & \ddots & & \\ & & & \mathrm{var}(\hat{\beta}_p) \end{pmatrix}$$

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2(X^tX)^{-1}_{jj})$$

BIAS-VARIANCE TRADE-OFF FOR STANDARD LINEAR MODELS

$$E\left[(y_0 - \hat{f}(x_0))^2\right] = \underset{x_i, y_0}{E}\left[(y_0 - x_0^t \hat{\beta})^2\right]$$

$$= \underbrace{\left[f(x_0) - E\left[\hat{f}(x_0)\right]\right]^2}_{\text{BIAS}} + \underbrace{\text{Var}(\varepsilon)}_{\sigma^2 \text{ irreducible error}} + \underbrace{\text{Var}\left(\hat{f}(x_0)\right)}_{\text{VARIANCE}}$$

BIAS $\quad x_0^t \beta - E(x_0^t \hat{\beta}) = x_0^t \beta - x_0^t E(\hat{\beta}) = x_0^t \beta - x_0^t \beta = 0$

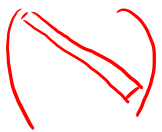
ZERO BIAS

VARIANCE $\quad \text{Var}(x_0^t \hat{\beta}) = x_0^t \, \text{Var}(\hat{\beta}) \, x_0 = \sigma^2 x_0^t (X^t X)^{-1} x_0$

$$= \sigma^2 \, \text{Trace}\left(x_0^t (X^t X)^{-1} x_0\right)$$

$$= \sigma^2 \, \text{Trace}\left(\underbrace{(X^t X)^{-1}}_{p \times p} \underbrace{x_0 x_0^t}_{p \times p}\right)$$

Trace of a Matrix

$A =$ 

sum of diagonal elements

$\text{Tr}(ABC) = \text{Tr}(BCA)$

$\text{cov}(X,Y) = E\left((X - E(X)) \cdot (Y - E(Y))\right)$

$= E(XY)$

$E(X) = E(Y) = 0$

$\widehat{\text{cov}}(X,Y) = \dfrac{\sum_{i=1}^{n} x_i y_i}{n}$

$\begin{pmatrix} x_{0i}^2 & x_{0i}x_{0i} & \cdots & x_{0i}x_{0p} \\ & & & \end{pmatrix}$

sample covariance

$$= \sigma^2 \, \text{Trace}\left((X^t X)^{-1} x_0 x_0^t \right)$$

$$= \sigma^2 \, \text{Trace}\left(\left(\underbrace{\frac{X^t X}{n}}_{I} \cdot n \right)^{-1} \underbrace{x_0 x_0^t}_{I} \right)$$

$$= \sigma^2 \, \text{Trace}\left(\frac{I}{n} \right)$$

$$= \begin{pmatrix} \frac{1}{n} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n} \end{pmatrix}_{p \times p}$$

$$\underbrace{\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}}_{p}$$

$$= \sigma^2 \frac{p}{n}$$

$$E\left[(y_0 - x_0^t \hat{\beta})^2 \right] = \sigma^2 + \sigma^2 \frac{p}{n}$$

Assuming that predictors are standardized (zero mean and unit variance) and uncorrelated, then $x_0 x_0^t = I$

Similarly $\frac{X^t X}{n}$ is the sample covariance of $(X_1, \ldots, X_p)$ from data, so $\frac{X^t X}{n} \approx I_p$

VARIANCE OF STANDARDS LINEAR MODEL

It depends on $\frac{p}{n}$ We can distinguish 3 cases:

① $n >> p$

→ bias = 0, variance very small, so LS will do very well in these cases

② n not much larger than p

→ variability in LS fit → overfitting, poor predictions

③ $n < p$ $X^t X$ is not invertible

$n = 100$
$p = 3000$

→ no unique LS estimate, infinite variance

There are 3 approaches to improve predictive accuracy of a standard linear model:

(1) Limit p ~~~ VARIABLE SELECTION

→ Reduces variance

→ Increases interpretability by identifying the most important predictors and removing redundant ones

(2) Shrinkage methods : Replace LS by alternative methods so that bias is increased (a bit) and variance is reduced

→ next lecture

(3) Dimensionality Reduction Project the p predictors into a M-dimensional space with $M < p$ and use the projections in a LS model

→ end of course

SUBSET SELECTION

Identify a subset of important predictors out of $X_1, \ldots, X_p$

Two procedures: best subset and stepwise selection

BEST SUBSET

Consider all possible combinations of the p predictors, fit a model by LS to each of those and choose best one according to a model selection criterion.

Algorithm ① Let $\mathcal{M}_0$ denote the null model (only intercept)
→ each observation is predicted to $\bar{y}$

② For $k = 1, \ldots, p$:

(a) Fit all $\binom{p}{k}$ models with exactly k predictors

(b) Pick the best among those and call this $\mathcal{M}_k$
↑ via RSS, R^2

③ Select a single best model out of $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ using some model selection criterion: AIC, BIC, C_p, adj R^2, CV

HOW MANY MODELS?

p predictors → 2^p possible models

p = 2 →

- $\mathcal{M}_0$ only intercept (null model)
- X_1
- X_2
- $\mathcal{M}_2 = X_1, X_2$ (full model)

} 4 models

Too many ↓ 2^p | Much less! ↓ $1 + \frac{p(p+1)}{2}$

p	2^p	$1 + \frac{p(p+1)}{2}$
2	4	
3	8	
8	256	37
16	65536	
20	1048576	211
100	1.26765×10^{30}	5051

STEPWISE METHODS

3 versions: FORWARD, BACKWARD, STEPWISE (hybrid)

FORWARD ① Let $\mathcal{M}_0$ be the null model

② For $k = 0, \dots, p-1$:

(a) Fit all $p-k$ models that augment $\mathcal{M}_k$ with one predictor

(b) Choose best among the $p-k$ models and call $\mathcal{M}_{k+1}$

③ Select best model among $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$ according to a model selection criterion

How many models now?

Null model → 1

$k=0$ → p models

$k=1$ → $p-1$ models

⋮

$k=p-1$ → 1 model

In total: $1 + \sum_{k=0}^{p-1} (p-k) = 1 + \sum_{k=1}^{p} (p-k+1) = 1 + p^2 + p - \sum_{k=1}^{p} k$

$= 1 + p^2 + p - \frac{p(p+1)}{2} = 1 + \frac{p(p+1)}{2}$

BACKWARD

- Starting from the full model (if you can!)
- Remove the least important predictor each time.

The main issue with forward or backward procedures is that once a predictor is added / removed from the model, then it remains so for all the subsequent steps. The problem is particularly pronounced for correlated predictors.

$\mathcal{M}_1$ X_2

$\mathcal{M}_2$ X_2, X_3

Hibrid approaches can be used here: STEPWISE

At each step of a forward procedure, one variable is added and any of the existing variables can be removed if it does not provide an improvement to the model fit.