Model Comparison + Information Criteria

Mallow's C_p
bias/variance tradeoff
prediction error

Prediction Problem: Given data $\underset{\sim}{y}$, $X = \begin{pmatrix} x_1 & \cdots & x_p \\ | & & | \end{pmatrix}$

1) Obtain OLS estimate $\hat{\underset{\sim}{\beta}}$

2) Construct fitted values $\hat{\underset{\sim}{\mu}} = X\hat{\beta} \in \mathbb{R}^{n\times 1}$

3) Use $\hat{\underset{\sim}{\mu}}$ to predict $\underset{\sim}{y}^*$, a new data vector generated under the same conditions as $\underset{\sim}{y}$, the original data vector.

this means $E[\underset{\sim}{y}] = E[\underset{\sim}{y}^*] = \underset{\sim}{\mu}$, (might not be $X\beta$!)

Q: How well does $\hat{\mu}$ do at predicting y^*?

A: Evaluate with expected prediction sum of squares

$$PSS = \| y^* - \hat{\mu} \|^2, \quad \text{evaluate } E[PSS]$$

Setup: $\underset{\sim}{y} = \underset{\sim}{\mu} + \underset{\sim}{\epsilon}$ $\quad V(\epsilon) = V(\epsilon^*) = \sigma^2 I$, ϵ, ϵ^* uncorrelated.

$\underset{\sim}{y}^* = \underset{\sim}{\mu} + \underset{\sim}{\epsilon}^*$

Given X, $\quad \hat{\beta} = (X^TX)^{-1}X^Ty$

$\hat{\mu} = X\hat{\beta} = X(X^TX)^{-1}X^Ty \equiv Py$ (P is a projection matrix)

IDENTITY #1: $E[\|y^* - \hat{u}\|^2] = n\sigma^2 + p\sigma^2 + \|(I-P)u\|^2$

sampling variability ↑ estimation var ↑ bias²

Proof: $\|y^* - \hat{u}\|^2 = \|u + \epsilon^* - \hat{u}\|^2$

$= \|\epsilon^* + (u - \hat{u})\|^2$

$= \|\epsilon^*\|^2 + \|u - \hat{u}\|^2 + \epsilon^{*T}(u - \hat{u})$

$E[\,''\,] = E[\|\epsilon^*\|^2] + E[\|u - \hat{u}\|^2] + E[\epsilon^{*T}(u - \hat{u})]$

$= n\sigma^2 + E[\|u - \hat{u}\|^2] + 0$

$= n\sigma^2 + E[\|u - \hat{u}\|^2]$

$\|u - \hat{u}\|^2 = \|u - Py\|^2 = \|u - Pu - P\epsilon\|^2$

$= \|(I-P)u - P\epsilon\|^2$

$= \|(I-P)u\|^2 + \|P\epsilon\|^2 - 2\epsilon^T P(I-P)u$

$= \|(I-P)u\|^2 + \epsilon^T P\epsilon - 0$

$E(\,'')\quad = \|(I-P)u\|^2 + p\sigma^2$

So $E[\|y^* - \hat{u}\|^2] = n\sigma^2 + p\sigma^2 + \|(I-P)u\|^2$

① All else being equal, prediction error is increasing in p

② Suppose $u = X\beta$. Then $(I - P_x)u = X\beta - X(X^TX)^{-1}X^TX\beta$

$= 0$ Bias is zero.

③ Suppose $X = [X_1, X_2]$ but $\hat{u} = X_1\beta$ (too many predictors)

$X_1 \in \mathbb{R}^{n \times p_1}$, $X_2 \in \mathbb{R}^{n \times p_2}$

if we use $\hat{u} = X_1 (X_1^T X_1)^{-1} X_1^T y$, then

$= P_1 y$

$E[\|y^* - \hat{u}\|^2] = n\sigma^2 + p_1\sigma^2 + \|(I - P_1)u\|^2$ ($\|(I-P_1)u\|^2 \to 0$)

$= n\sigma^2 + p_1\sigma^2$

if we use $\hat{u} = X(X^TX)^{-1}X^Ty$ ($X = (X_1, X_2)$, too many predictors)

$= P_X y$

$E(\|y^* - \hat{u}\|^2) = n\sigma^2 + (p_1 + p_2)\sigma^2 + \|(I - P_X)u\|^2$ ($\|(I-P_X)u\|^2 \to 0$)

$= n\sigma^2 + (p_1 + p_2)\sigma^2$

So adding unnecessary predictors increases the prediction error.

(4) If $u = [X \; X^*]\beta = \tilde{X}\beta$ (not enough predictors)

then $E(\|y^* - \hat{u}\|^2) = n\sigma^2 + np^2 + \|(I - P_X)u\|^2$

$\|(I-P_X)u\|^2$ not zero.

In general, $E[\|y^* - \hat{u}\|^2] = n\sigma^2 + p\sigma^2 + \|(I-P)u\|^2$

- $p\sigma^2$: goes up when adding predictors
- $\|(I-P)u\|^2$: goes down when adding predictors.

Goal: Add predictors that reduce bias
Don't add unnecessary predictors

How to do this? Find a way to estimate the prediction error

IDENTITY #2: $E[\|y^* - \hat{\mu}\|^2] = E[RSS] + 2p\sigma^2$

This means $RSS + 2p\sigma^2$ is an unbiased est. of $E[PSS]$

In particular, this shows that 1) RSS is a bad est. of PSS

2) RSS gets worse at est PSS as p gets bigger.

In general, dont know σ^2; but suppose $\hat{\sigma}^2$ is an unbiased est.

Then $RSS + 2p\hat{\sigma}^2$ is an unbiased est. of $E[PSS]$.

Model Selection with C_p

Consider two models, for y, $y \sim X_1$ $\text{ncol}(X_1) = p_1$ Model 1
$y \sim X_2$ $\text{ncol}(X_2) = p_2$ Model 2

1) fit M1, get RSS_1
2) fit M2, get RSS_2
3) obtain estimate $\hat{\sigma}^2$ of σ^2 by fitting a model with all available regressors.

Then, prefer M1 to M2 if $RSS_1 + 2p_1\hat{\sigma}^2 < RSS_2 + 2p_2\hat{\sigma}^2$

$RSS_j + 2p_j\hat{\sigma}^2$ is the "C_p statistic" for model j.

$C_p(M_j) = RSS_j + 2p_j\hat{\sigma}^2$

Sometimes, C_p is defined as $C_p(M_j) = \frac{RSS_j}{\hat{\sigma}^2} + 2p_j - n$

but this doesnt matter for model comparison.