6. Linear Model Selection & Regularization.

• **Prediction accuracy:**
- least squres work well if:
  - true relationship between the response & predictors is approximately linear
  - $n >> p$ (works well on test)
- if $p > n$ → no unique solut^n to $\hat{\beta}$.
  → greater variability ⇒ poor test performance
- constraining / shrinking the coefficients as solution to this problem.

Model interpretability:
- removing irrelevant predictors.
- excluding irrelevant variables from MLR: feature selection / variable selection.

(1) Subset selection:
- identifying a subset of the p predictors that we believe to be related to the response. → followed by least squares.

2) Shrinkage / Regularization:
- shrinking estimated coefficients towards zero relative to least square estimates.
- to reduce variance
- a few coeff can be 0 ⇒ also used for feature selection.

(3) Dimension reduction
- projecting the 'p' predictors into an M-dimensional subspace $(M<p)$.
- achieved by computing M different linear combinations / projections of the variables.
- then these M projections are used as predictors to fit a linear regression by least squares.

6.1 Subset selection.

6.1.1. Best Subset selection.

* why other estimates than R^2?
→ RSS of p+1 models decreases monotonically
⇒ R^2 ↑ monotonically
⇒ no of features included in the models ↑ ⇒ full model always selected.

high R^2: low train error, we need low test error.

M_0 :: the null model (only intercept)

↓

For $k = 1(1)p$
fit all $\binom{p}{k}$ and choose the one with smallest RSS or highest R^2 (M_k).

↓

* Select single best model from $M_0, M_1, \ldots, M_p$ using C_p/AIC/BIC/adjusted R^2/ predict^n error on validat^n set / cross validation method.

* Different criteria for evaluating models.

(i) coefficient of multiple regression.

Let there be p many predictors in the model.

$$R^2_p = \frac{SS_{reg}(p)}{SST} = 1 - \frac{SS_{res}(p)}{SST}.$$

(ii) adjusted $R^2_p = \bar{R}^2_p$

$$\bar{R}^2_p = 1 - \frac{MS_{Res}(p)}{MS_T} = 1 - \frac{SS_{Res}(p)}{n-p} \frac{n-1}{SS_T}$$

$$= 1 - \frac{n-1}{n-p} \cdot \frac{SS_{Res}(p)}{SST}$$

$$= 1 - \frac{n-1}{n-p} (1 - R^2_p)$$

(iii) Mallow's C_p: It measures the overall bias or mean square error in the fitted model.

* low C_p and $C_p \approx p$ indicates it is best model,

$$C_p = \frac{SS_{res}(p)}{MS_{res\ Full}} - n + 2p.$$

$$C_p = MSE + \underbrace{(\text{penalty})}_{\text{estimate of bias/ cost on models for having extra parameters.}}$$

(iv) Akaike Information Criterion: (AIC)

$$AIC = -2\log(L) + 2k.$$

k: no. of parameters.
L: maximized likelihood of model

goal: to select best model for prediction.

(v) Bayesian Information Criterion: (BIC)

$$BIC = k\ln(n) - 2\ln(L)$$

k: no of parameters.
n: no of samples
L: maximized likelihood of model

goal: to select best model for explanation.

(vi) Deviance = $-2\ln(L)$

- smaller the deviance better the fit.

- best subset selection becomes computationally infeasible for larger values of p.

6.1.2 Stepwise selection.

① Forward Stepwise.

1) M_0: null model : only intercept.

2) For $k = 0, \ldots, p-1$

(a) consider all $p-k$ models that augment the predictors in M_k with one additional predictor

(b) M_{k+1} = the best among $p-k$ models (RSS or R_p^2)

3) select single best model from $M_0, M_1, \ldots, M_p$ using C_p/AIC/BIC/adj R_p^2/CV etc.

* Total no. of models to be fitted:

$$1 + \sum_{k=0}^{p-1}(p-k) = \underbrace{1 + \frac{p(p+1)}{2}}$$

Substantially less than 2^p

* ~~may~~ not guaranteed to find out the best possible model in 2^p.

* if $n<p$: $M_0, \ldots, M_{n-1}$ total possible fits as for $p \geq n$, not unique solution possible.

② Backward Stepwise.

1) M_p: full model : all p predictors.

2) for $k = p, p-1, \ldots, 1$;

(a) Consider all k models that contain all but one of the predictors in M_k for a total of $k-1$ predictors;

(b) choose the best among these k : M_{k-1} (RSS or R_p^2)

3) Select single best model from $M_0, \ldots, M_p$ using C_p/AIC/BIC/adj R_p^2/CV etc.

* $1 + \frac{p(p+1)}{2}$ total no. of models to be fitted.

* Possible only if $n > p$.

③ Hybrid approach.

forward selectⁿ along with removing variables which are irrelevant at every stage.

① Let d be no. of predictors in model

$$C_p = \frac{1}{n}(RSS + \underbrace{2d\hat{\sigma}^2}_{})$$ ~ Mallow's C_p : $C_p' = \frac{RSS}{\hat{\sigma}^2} + 2d - n$.

→ penalty : to adjust for the fact that the training error tends to underestimate the test error

When $\hat{\sigma}^2$ is UE of σ^2 then C_p is UE of MSE.

∴ Choose the model with lowest C_p.

② AIC criterion - defined for large class of models fit by MLE.

6.2. Shrinkage Methods.

6.2.1 Ridge regression.

$\hat{\beta}^R$: the ridge regression coefficients.

minimize:

$$\underbrace{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{RSS}} + \lambda \underbrace{\sum_{j=1}^{p}\beta_j^2}_{\text{shrinkage Penalty}}$$

tuning parameter: $\lambda \geq 0$.

(small when $\beta_1, \beta_2, \dots, \beta_p$ are close to zero).

- if $\lambda = 0 \Rightarrow$ least squares estimates.
- as $\lambda \to \infty$, the impact of the shrinkage penalty grows & the ridge regression coefficient estimates will approach zero.
- For each λ, ridge regression will produce $\hat{\beta}_\lambda^R$ for each value of λ.
- observe that, we want to shrink the estimated association of each variable with the response; however we do not want to shrink the intercept, which is simply a measure of <u>mean</u> <u>value</u> <u>of the</u> <u>response</u> when $x_{i1} = x_{i2} = \dots = x_{ip} = 0$.
- if X is centered to have mean 0, intercept for ridge will be

$$\hat{\beta}_0 = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

- standard least squares coefficient are scale equivariant: multiplying x_j by constant $c \Rightarrow \hat{\beta}_j$ gets scaled by $1/c$.

$\Rightarrow$ regardless of how the j^{th} predictor is scaled $x_j\hat{\beta}_j$ will remain the same.

- ridge regression estimates can change substantially when multiplying a given predictor by a constant.
- It is best apply ridge regression after standardizing the predictors:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

*why does Ridge Regression improve over least squares?

- bias variance trade off
- substantial computational advantage.

6.2.2. The Lasso Regression.

- lasso is a relatively recent alternative to ridge regression that overcomes this disadvantage of variable selection.

minimize:

$$\underbrace{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{RSS} + \overbrace{\lambda}^{\text{tuning parameter}} \underbrace{\sum_{j=1}^{p}|\beta_j|}_{\text{penalty}}$$

(has the effect of forcing some of the coefficient estimates to be exactly equal to zero when λ is sufficiently large)

- lasso performs variable selection.
- lasso yields sparse models - that is models that involve only a subset of the variable.
- ~~Ridge~~ Lasso regression can also be represented as:

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \text{ subject to } \sum_{j=1}^{p}|\beta_j| \le s$$

Ridge regression:

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \text{ subject to } \sum_{j=1}^{p}\beta_j^2 \le s$$

- we perform the lasso we are trying to find the set of coefficient estimates that lead to the smallest RSS, subject to the constraint that there is a budget s for how large $\sum_{j=1}^{p}|\beta_j|$ can be.

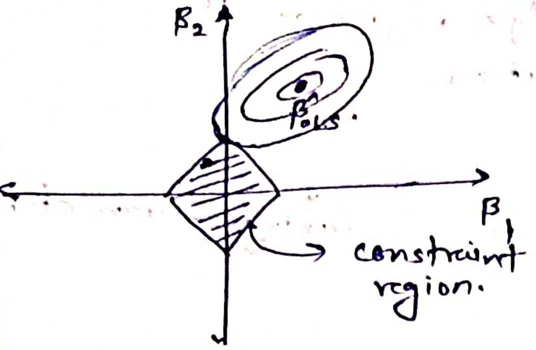
Subset selection:

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \text{ subject to } \sum_{j=1}^{p} I(\beta_j \neq 0) \le s.$$

i.e. finding a set of coefficient estimates such that RSS is as small as possible, subject to constraint no more than s coefficients can be non-zero. $\longrightarrow$ this is equivalent to best subset selection.

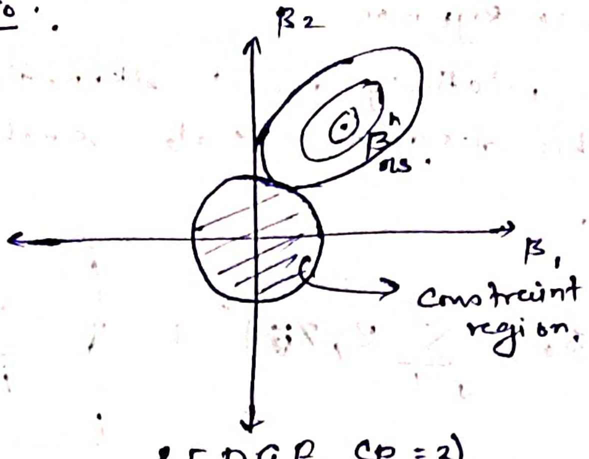
- lasso performs feature selection for s sufficiently small.

* The variable selection property of Lasso.



LASSO (p=2)

intersection is on axis
⇒ LASSO performing variable selection.



RIDGE (p=2)

intersection is usually not on axis.
⇒ RIDGE cannot perform variable selection.

- lasso leads to feature selection when $p > 2$ due to the sharp corners of polyhedron or polytope.

* comparing Lasso & the ridge regression.

- the lasso implicitly assumes that a number of the coefficients truly equal to zero.
- lasso performs well compared to ridge, when ~~when~~ only a few parameters are truly related to response, else ridge will perform better.
- ridge will perform better when the response is function of many predictors all with coefficients of roughly equal size.
- lasso solution can yield a reduction in variance at the expense of a small increase in bias ⇒ more accurate prediction.
- lasso performs variable selection hence results in models that are easier to interprete.

* Special case:

- $n = p$,
- X a diagonal matrix with 1's on the diagonal and 0's in all ~~diag~~ off diagonal elements.
- regression without intercept. : $\beta_1, \ldots, \beta_p$.

Least square: $\sum_{j=1}^{p} (y_j - \beta_j)^2 \xrightarrow{\text{minimize}} \hat{\beta}_j = y_j$

Ridge: $\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \xrightarrow{\text{minimize}} \hat{\beta}_j^R = y_j / (1+\lambda)$

Lasso: $\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \xrightarrow{\text{minimize}} \hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2; \\ 0 & \text{if } |y_j| \le \lambda/2 \end{cases}$

"soft thresholding"

in summary.

- ridge regression: more or less shrinks every dimension of the data by same proportion.
- lasso regression more or less shrinks all coefficients toward zero by a similar amount & sufficiently small coefficients are shrunken all the way to zero.

* Bayesian Interpretation of Ridge Regression & Lasso.

- we can view ridge and Lasso through a Bayesian Lense.
- Bayesian viewpoint for regression assumes that the coefficient vector β, has some prior distribution $p(\beta)$, $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$.

By Bayes' thm: $p(\beta|X,Y) \propto f(Y|X,\beta)\, p(\beta|X) = f(Y|X,\beta)\cdot p(\beta)$ & assuming X is fixed

∴ Assume usual linear model ① $Y = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p + \epsilon$.
② $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ ③ $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$

→ If $g: N(0, f(\lambda))$ then MLE of β is ridge regression estimates of beta.

→ If $g: L(0, f(\lambda))$ then MLE of β is Lasso estimates of beta.

* Proof:

① ~~Ridge~~ LASSO.

Let $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Likelihood of data:

$$L(Y|X,\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2\right)$$

$$p(\beta) = \frac{1}{2b} \exp\left(-\frac{|\beta|}{b}\right) \qquad (\lambda = 1/b).$$

As $p(\beta|X,Y) \propto L(Y|X,\beta)\, p(\beta|X) = L(Y|X,\beta)\, p(\beta)$

$$\therefore L(Y|X,\beta)\cdot p(\beta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \frac{1}{2b} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2 - \frac{|\beta|}{b}\right)$$

$$\therefore \ln L = -n\ln\sigma\sqrt{2\pi} - \ln 2b - \frac{\sum_{i=1}^{n}\epsilon_i^2}{2\sigma^2} - \frac{|\beta|}{b}.$$

$$\text{maximize } \ln L = \underset{\beta}{\text{minimize}} \frac{\sum_{i=1}^{n}\epsilon_i^2}{2\sigma^2} + \lambda\sum_{j=1}^{p}|\beta_j| \longrightarrow \text{LASSO!}$$

(2) Ridge. $\beta_i \overset{iid}{\sim} N(0, c)$

$$p(\beta) = \prod_{i=1}^{p} p(\beta_i) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi c}} \exp\left(-\frac{\beta_i^2}{c}\right) = \left(\frac{1}{\sqrt{2\pi c}}\right)^p \exp\left(-\frac{1}{2c} \sum_{i=1}^{p} \beta_i^2\right)$$

$\therefore$ ~~β~~ $p(\beta|X,Y) \propto \mathcal{L}(Y|X,\beta) p(\beta|X) = \mathcal{L}(Y|X,\beta) p(\beta)$.

$\therefore L(Y|X,\beta) p(\beta)$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \left(\frac{1}{\sqrt{2\pi c}}\right)^p \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 - \frac{1}{2c} \sum_{i=1}^{p} \beta_i^2\right).$$

$$\underset{\beta}{\text{maximize}} \log\left(L(Y|X,\beta) \cdot p(\beta)\right)$$

$$= \underset{\beta}{\text{minimize}} \frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 + \frac{1}{2c} \sum_{i=1}^{p} \beta_i^2 \longrightarrow \text{Ridge!!}$$

- select ~~t~~tunning parameter by cross validation ~~the~~ s.t. test MSE is least.