Linear model selection & Regularization

in the regression setting, the standard linear model

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

in real world problems, linear models have advantages such as interpretability and prediction accuracy.

prediction accuracy - if the true relationship is linear, least squares will have ~~test error~~ low bias

if $n \gg p$, linear model will have less variance

if n is not much larger than p, then it will variability & that model will overfit

if $p > n$, then there's no single values for coefficients, it will perform poorly on test set

by constraining or shrinking, we can substantially reduce the variance at the cost of ~~bias~~ negligible bias

model interpretability - including irrelevant variables lead to unnecessary complexity in the resulting model

by ~~setting~~ corresponding coefficients to zero, we obtain a model that is more easily interpreted.

Some classes of methods to fit least squares

- subset selection: identify subset of the p predictors that we believe to be related to the response
- shrinkage: fitting the model with all predictors, then estimated coeff are shrunken towards zero. this shrinkage or regularization has the effect of reducing variance.
- dimension reduction: project the p predictors into M dimensional subspace, where $M < p$. then these M projections are used as predictors to fit linear regression by least squares. projections are calculated by M different linear combinations

Subset selection

fit p predictors into 2^p models. **best subset selection**

null model, models with only 1 predictor, models with 2 p, ----

upto models with all p

in each step find the best one - that gives smallest RSS or largest R^2

finally you are left with $M_0, M_1, \ldots, M_p$

now select from these p+1 models using prediction error on a validation set, C_p (AIC), BIC or adjusted R^2.

or use cross validation.

R^2 increases as we add new predictors.

find a right balance of min RSS & high R^2.

drawback of subset selection 2^p models are computationally intense

in case of logistic regression, we use deviance instead of RSS.

deviance = $-2 \times$ max log likelihood

smaller the deviance, better fit.

stepwise selection

forward stepwise selection: add predictors one at a time, keep the best one & proceed. find the best one from M_p. not guaranteed to select the best combination of predictors.

works even when $n < p$. total models - $1 + p(p+1)/2$

backward stepwise selection: fit all predictors. chose the one thats best. next fit ~~all but one~~ p-1 predictors. select the best. then fit p-2. select the best. so on. finally choose from these best models

n should be greater than p. total models - $1 + p(p+1)/2$

hybrid - add new variables. but also remove if the added variable is not providing value

Choosing the optimal model

training error is not a good way to select the best model
test error should be the criteria

2 methods - adjust the training error for bias, overfitting
- validation set approach / cross validation approach

$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$

d - no of predictors

$\hat{\sigma}$ - estimate of variance using the model with all predictors

C_p statistic adds a penalty of $2d\hat{\sigma}^2$ to the training error

(Mallow's $C_p = \cancel{\frac{RSS}{\hat{\sigma}^2 + 2d - n}} \quad \frac{RSS}{\hat{\sigma}^2} + 2d - n$)

penalty increases as no. of predictors increase

in case of linear regression model with Gaussian errors, max likelihood and least squares are the same thing.

$$AIC = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

BIC is derived by Bayesian point of view

$$BIC = \frac{1}{n}(RSS + \log(n)\, d\, \hat{\sigma}^2)$$

C_p, AIC, BIC - smaller value → smaller test error

Adjusted R^2 - higher value → smaller test error

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

unlike the R^2 statistic, adjusted R^2 pays price for inclusion of unnecessary variables

CP, AIC, BIC are more preferred statistically than adjusted R^2.

AIC, BIC can be used for more general types of models as well.

validation & cross validation - rather than predicting the test error, here we actually know it.

the errors would differ based on the split or folds.

in case of subset selection, the validation errors are averaged over all folds for each model size k.

'one standard error rule' - we first calculate the std error of estimated MSE for each model size. then select the ~~model~~ smallest model for which the estimated error is within 1 std error of lowest point on the curve.

Shrinkage methods

Ridge regression

$\hat{\beta}^R$ are the coeff that minimize

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

where $\lambda \geq 0$ - tuning parameter

$\lambda \sum_j \beta_j^2$ - shrinkage penalty

when $\lambda = 0$, penalty term has no effect. they are same as least square estimates

when $\lambda \to \infty$, penalty increases & ridge coefficients approach 0. (but never zero)

β_2 L2 norm $\quad \|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

when $\lambda = 0$, variance is high, bias is 0

as λ increase variance decreases with a slight increase in bias

apply ridge regression after standardizing the predictors

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

ridge regression works well when least square estimates have high variance, when p>n,

Lasso

disadvantage in ridge - includes all variables & then shrinks them towards zero. interpretability is difficult for large p.

lasso coeff tries to minimize the quantity

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

l_1 norm coefficient

$\|\beta\|_1 = \sum |\beta_j|$

when $\lambda = 0$, lasso coeff are same as least square estimates
when λ increases, some of the coeff might become zero
it will act as variable selection. and lasso be called as sparse model

as with ridge regression, selecting a good value of λ is important.

when $p=2$, (s - no. of predictors in subset)
lasso coefficients have lowest RSS out of all points that lie within the diamond defined by $|\beta_1| + |\beta_2| \le s$
ridge coeff - points within the circle $\beta_1^2 + \beta_2^2 \le s$

lasso will perform better when small no. of predictors have substantial coefficients. others could be very min or zero; ridge will be better when all predictors are essential. choosing how? cross validation could help.

r^v to ridge, lasso will decrease the variance with a small increase in bias.

w/o intercepts, we can show that for the min co efficients

$$\hat{\beta}_j^R = y_j / (1+\lambda)$$

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2 \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2 \\ 0 & \text{if } |y_j| \leq \lambda/2 \end{cases}$$

for a certain value, the coefficients in Lasso are zero.
" soft thresholding"

Ridge regression shrinks every dimension of data by same proportion
Lasso shrinks towards zero by similar amount, sufficiently small coefficients are shrunken all the way to zero.

selecting the tuning parameter: cross validation on a chosen no. of values for λ. then select the one for which the cross validation error is smallest. finally use that λ, refit the model on all available observations. ~~for the selected~~

<u>Dimension Reduction methods</u>

transform the predictors & then fit a least squares model using the transformed variables.

let $Z_1\ Z_2 \dots Z_m$ represent $M<p$ linear combinations of our original p predictors,

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

for some constants $\phi_1\ \phi_2 \dots \phi_m$ $m=1,2,\dots,M$. we can then fit linear regression model

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i \qquad i=1,\dots,n$$

if ϕ are chosen correctly, then the above regression will outperform least square regression.

The problem of estimating p+1 coeff becomes reduced to M+1
"dimension reduction"

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$$

dimension reduction constrains the estimated β_j coefficients, has the potential to bias the coeff estimates.
when $M << p$, reduces the variance.
if $M = p$, its equivalent to least squares regression.
& all Z_m are linearly indepenent

steps - transformed predictors $Z_1 Z_2 . Z_m$ are obtained
model is fit using these predictors.
ϕ_{jm} can be obtained in different ways such as principal components and partial least squares

Principal components regression
PCR
reduces the dimension of an $n \times p$ data matrix X.
- the first principal component direction of the data is that along which the observations vary the most.
(the line which answers most variance of data)
- the first principal component vector defines the line that is as close as possible to the data
the points in the data should be at least distance from the line
- out of every linear combination of variables, $\phi_{11}^2 + \phi_{12}^2 = 1$ yields the highest variance
it is necessary to consider only linear combinations of the form $\phi_{11}^2 + \phi_{12}^2 = 1$, otherwise any arbitrary values could be chosen to blow up the variance
- ex: $Z_{i1} = 0.839 \times (pop_i - \overline{pop}) + 0.544 (ad_i - \overline{ad})$
the values $Z_{11}, Z_{21}, \ldots, Z_{n1}$ are principal component scores

- the second principal component direction must be perpendicular to the 1st pc direction. i.e uncorrelated
- subsequent additional components will maximize the variance, subject to the constraint being uncorrelated with the preceding components

ex:- . 2nd PCA for the prev example

$$Z_2 = 0.544 \times (pop - \bar{pop}) - 0.839 \, (ad - \bar{ad}).$$

if both the variables are linear & captured by z_1 itself & z_2 provides very less variance, then z_1 is sufficient.
plot of ~~variable~~ z_1 scores vs variable will explain the relationship

if $M << p$, then PCR will perform better as it will avoid overfitting
also PCR will be better only if majority of variance is captured by initial five components, else ridge & lasso would be better.
PCR is not a feature selection, because the components are constitute of individual predictors, but their projections.
so we can say PCR is more $\parallel^{r}$ to Ridge rather than Lasso

before performing PCR, all variables should be on the same scale.
else standardize them via

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

Partial least squares

drawback of PCR - there's no guarantee that the directions that best explains the predictors will also be the best directions to use for predicting the response (unsupervised)

unlike PCR, PLS identifies these new features in a supervised manner, new features not only approximate the old ones, but also related to the response

- PLS first computes Z_1 by setting each ϕ_{j1} equal to the coeff from simple linear regression of Y onto X_j
  hence in computing $Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$, PLS places highest weight on the variables that are most strongly related to the response.

- PLS direction doesn't fit the predictors as closely as PCA. but does a better job at explaining the response
- to identify Z_2, we take the residuals after regressing each variable on Z_1. then compute Z_2 using this orthogonalized data ııı^{r} to how we computed Z_1
- finally fit a least square model using these M predictors
- M - no. of partial least squares is found out using cross validation
- while PLS reduces bias, it increases variance as well. so the overall effect of PLS is not effective to PCR.