---

02. Linear Model

least-square approach minimises $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

if $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$.

coefficient accuracy: $SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_i (x_i-\bar{x})^2}$, $SE(\hat{\beta}_0)^2 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i-\bar{x})^2}\right)$ $\Rightarrow$ $t = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{SE(\hat{\beta}_1)} \sim t(n-2)$

overall fitness: (1) $RSE = \sqrt{\frac{1}{n-2}RSS}$

(2) $R^2 = \frac{TSS - RSS}{TSS}$ → fraction of variance explained (how much does the model reduce TSS)

where $TSS = \sum_i (y_i - \bar{y})^2$, $RSS = \sum_i (y_i - \hat{y}_i)^2$.

for a simple regression model: $r = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i (x_i-\bar{x})^2}\sqrt{\sum_i (y_i-\bar{y})^2}}$

hierarchy principle: If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

How to allow qualitative predictors: by creating a dummy variable, i.e. create a new variable s.t. (with no dummy variables → baseline)

$x_{i1} = \begin{cases} 1 & \text{if ith observation belongs to A} \\ 0 & \text{if ith observation does not belong to A} \end{cases}$

$x_{i2} = \begin{cases} 1 & \text{if ith observation belongs to B} \\ 0 & \text{if ith observation does not belong to B} \end{cases}$

$\Rightarrow y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots = \begin{cases} \beta_0 + \varepsilon_i & \\ \beta_0 + \beta_1 x_{i1} + \varepsilon_i \dots A & \\ \beta_0 + \beta_2 x_{i2} + \varepsilon_i \dots B & \end{cases}$

Is at least one predictor useful?: $F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$

which predictors are useful?: subset selection

which model (different combination of predictors) is the best: model selection.

03 Linear Model Selection and Regularization

1. Linear Model Selection: subset selection (R^2, RSS) + model selection (AIC, BIC, C_p, adj. R^2, CV)

1.1 Subset Selection

1.1.1 Exhausted / best subset

method: consider all possible models; and compute least square fit for all of them, then choose the best model based on some criterion.

steps:

(1) Let M_0 be the **null** model, which contains only the intercept $\bar{y}$.

(2) For $k = 1, 2, \dots p$.

(a) fit all $\binom{p}{k}$ models

(b) Pick the **best** (smallest RSS, or largest R^2) model and let it be M_k.

(3) Compare all p models and select the one with the least **test** error (→ model selection)

problem: We need to consider all $1 + \binom{p}{1} + \binom{p}{2} + \dots + \binom{p}{p} = 2^p$ models, if $p = 40$, 2^p = 1 trillion.
Also the problem of overfitting and high variance ⇒ stepwise selection.

1.1.2 Forward Selection

method: start with null model, then pick the next predictor that reduces RSS the most; put that predictor in the model, repeat until some stopping rule is satisfied.

steps:

(1) Let M_0 be the **null** model

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

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

(b) pick the **best** model among all $p-k$ models and let it be M_{k+1}.

(3) Compare all p models and select the one with the least **test** error

note: greater computing advantages since only $p + (p-1) + \dots + 1 + 1 \approx p^2$ (M_0) are considered.

problem: it's not guaranteed the forward selection would find the best combination of predictors. especially if there is correlation between predictors.

1.1.3 Backward Selection:

method: start with full model, the remove the predictor with the largest p-value; then fit the model with $p-1$ predictors, repeat until some stopping rule is satisfied.

steps:

(1) Let M_p be the **full** model, which contains all p predictors.

(2) For $k = p, p-1, \dots 1$.

(a) Consider all k models that contain all but one predictors in M_k.

(b) pick the **best** model among all k models and let it be M_{k-1}.

(3) Compare all p models and select the one with the least **test** error

note: backwards selection requires $n > p$, whereas forward selection can be used even when $p > n$

problem: same as forward.

1.2 Model Selection

To choose an 'optimal' member in the path of models
(i.e. # predictors & which predictors).

Why test error? - The full model will always have the lowest training error, i.e. overfitting
- Training error is not a represent of how good the model predict on new data.

how? ① by adjusting training error to account for overfitting; ② by seperating test/train data.

adjust RSS

1. Mallow's C_p (C_p)

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

(d → # predictors, e.g. M_3 has 4 predictors including intercept.)

($\hat{\sigma}^2$ → estimate of $Var(\varepsilon)$, i.e. variance of error.)

problem: C_p must have $n > p$; if $n \leq p$, M_p is not defined and has $C_p = 0$.

2. Akaike Information Criterion (AIC).

$$AIC = -2\log L + 2d$$

(L → maximum likelihood.)

note: For linear model with Gaussian error, C_p = AIC because $-2\log L = \frac{RSS}{\hat{\sigma}^2}$.

i.e. $AIC = \frac{RSS}{\hat{\sigma}^2} + 2d \iff C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$

i.e. $\frac{nC_p}{\hat{\sigma}^2} = AIC$.

proof: For linear model with Gaussian errors. $\Rightarrow \varepsilon_i \sim N(0, \sigma^2) \Rightarrow y_i \sim N(\hat{y}_i, \sigma^2)$.

$\Rightarrow L = \exp\{-\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{2\sigma^2}\} = \exp\{-\frac{RSS}{2\sigma^2}\} \Rightarrow \log L = -\frac{RSS}{2\hat{\sigma}^2} \Rightarrow -2\log L = \frac{RSS}{\hat{\sigma}^2}$

problem: also requires $n > p$.

3. Bayesian Information Criterion (BIC).

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

($\log(n)$ → replace 2 with $\log(n)$ from C_p.)

note: BIC places a heavier penalty on size of model than C_p or AIC

adjust R^2 ← 4. adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \qquad [R^2 = \frac{TSS - RSS}{TSS} \text{ doesn't account for size of model}]$$

note:
- effective part $RSS/(n-d-1) \Rightarrow$ place a penalty on including unnecessary variables since as $d\uparrow$, $RSS/(n-d-1)$ may $\uparrow$ or $\downarrow$, and so does adjusted R^2
- adjusted R^2 doesn't require $n > p$.

seperate test & train data. ← 5. Cross-Validation (CV).

Ideally, we would split* the data into 3 sections for model selection and assessment.

* usually split into 2:1:1

model fitting

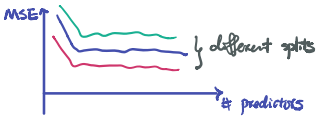
model selecting (e.g. select model from subsets select, or selecting λ for lasso)

model assessment (i.e. prediction accuracy).

As we focus on model selection now, we only need TRAIN and VALD sets for now. When we assess model performance, we'll use TEST set. (more in 05 GMD).

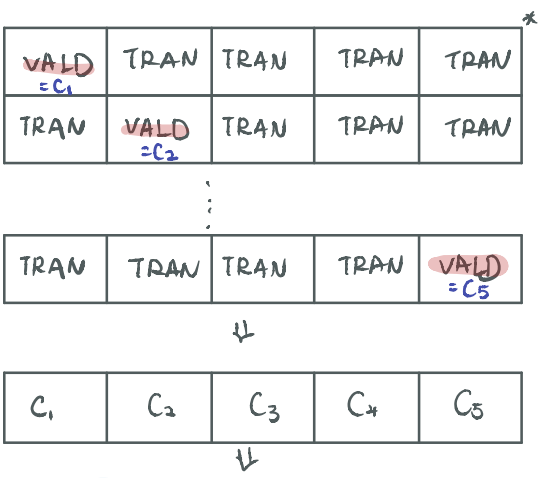
- Validation:

fitting the model. MSE / MCR.



problem:
- even though the validation estimated test error tell us which model has the lowest MSE, the validation estimated test error itself varies a lot. $\Rightarrow$ no actual level of test error.
- a fraction of data is not used to train the model $\Rightarrow$ waste lots of data.

- **Cross-validation:**



* 5-fold, i.e. $k=5$.

steps:
(1) randomly split the data into k parts.
(2) each part will be the validation part in turn, and the rest $k-1$ pieces are the training part
(3) fit the model using the training data and compute the error using the validation part, repeat k times
(4) compute the average error

CV error $= CV_{(k)} = \sum_{k=1}^{k} \frac{n_k}{n} MSE_k$ or $\sum_{k=1}^{k} \frac{n_k}{n} Err_k$.

where $MSE_k = \sum_{i \in C_k} \frac{(y_i - \hat{y}_i)^2}{n_k}$,

$Err_k = \sum_{i \in C_k} \frac{I(y_i \neq \hat{y}_i)}{n_k}$,

n_k = # observ. in k^{th} part

note: - letting $k = n$, yields n-fold or leave-one-out cross validation (LOOCV)
i.e. every data point gets to be training data as well as validation data.

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2$$, where h_i is the leverage (i.e. diagonal of the "hat" matrix).

- Generally k-fold preferrable on LOOCV.
  - computation time shorten.
  - LOOCV has high variance because each training data is highly similar/correlated
    Thus, even though LOOCV has low bias, the tradeoff between bias and variance is not worth it

Important: subset selection should be done within CV process! cannot select the predictors and then do CV!

* One-standard-error rule: If test error are very close for some model, we first calculate the standard error of the test error for each model, the select the smallest model for which the test error is within one standard error of the lowest point on the curve.

2. Shrinkage/Regularization (shrinking coefficients toward or to exactly 0)

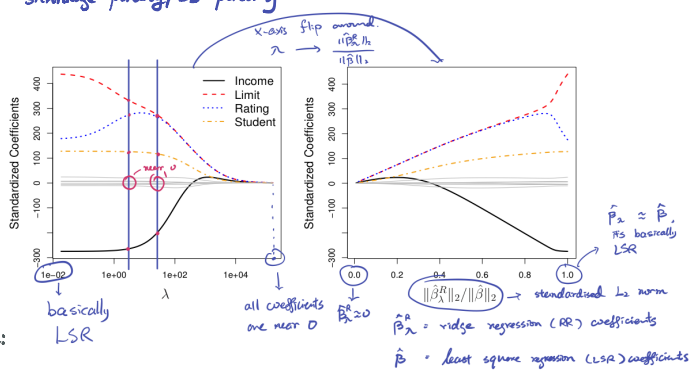
↳ Ridge ↳ Lasso.

2.1 Ridge Regression

minimize $\underbrace{\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{\text{shrinkage penalty / } L_2 \text{ penalty}}$ ← tuning parameter

- RSS term ensure fitness of data; shrinkage penalty has the effect of shrinking estimated coefficients towards 0. (but $\neq 0$).
- tuning parameter λ controls the relative impact of these two terms.
- choosing a good tuning parameter is critical; can use CV to determine.
- Since $\sum_{j=1}^{p} \beta_j^2$ is multiplied by a constant λ, the Ridge coefficient estimates can change quite dramatically. Thus, it's best to standardize first:

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



$\hat{\beta}_\lambda^R$ = ridge regression (RR) coefficients

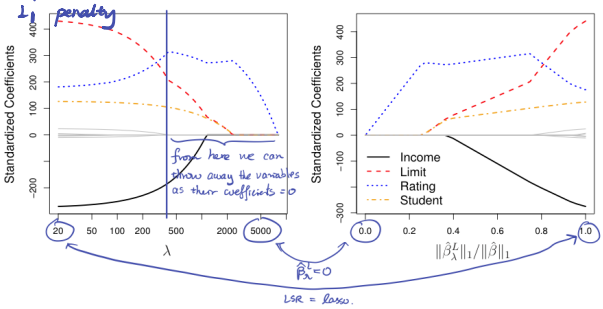
$\hat{\beta}$ = least square regression (LSR) coefficients

L_2 norm $= \|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$

2.2 **Lasso**

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

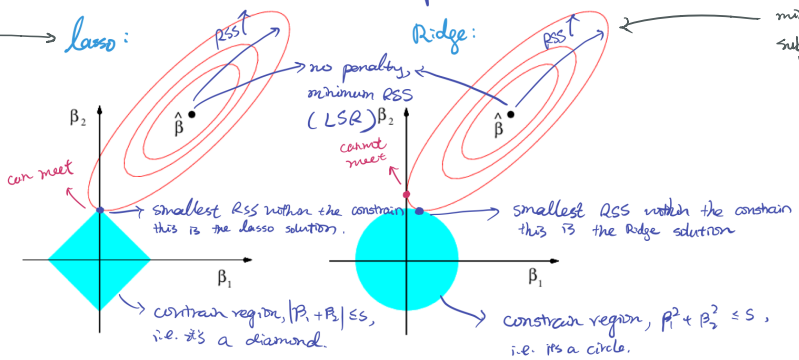
- l_1 penalty has the effect of forcing some of the coefficient estimates to be exactly 0 when λ is large enough
- Thus lasso can perform predictor selection like subset selection, i.e. it yields sparse model
- CV again is used to choose λ
- Why does lasso can force coefficient estimates to be 0 but not Ridge:



The Lasso Picture (2 parameters example)

minimize RSS subject to $l_1 \le s$ → **Lasso:**

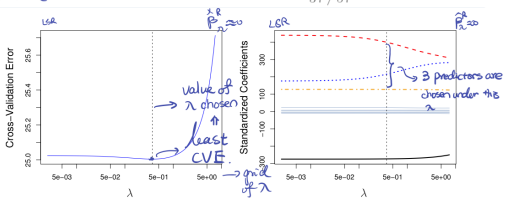
Ridge: ← so RSS, minimize to l_2 eg. subje



* the diamond has corners (for higher dimensions, we have edges and corners), so one parameter is 0 on y-axis and it's easier for RSS circle to hit it.

37 / 57

* How to choose λ using CV:
  1. choose a grid of λ
  2. compute CVE for each λ
  3. select the λ that has the lowest CVE.



3. Dimension reduction (linear combination of predictors, e.g. PCR, PLS).

method: transform the predictors into linear combination and then fit a least squares model

step: (new predictor, constant, original predictors)

(1) let $z_m = \sum_{j=1}^{p}\phi_{mj}x_j$

(2) fit the model: $y_i = \theta_0 + \sum_{m=1}^{M}\theta_m z_{im} + \varepsilon_i$, $i = 1, 2, \dots n$.

note:

- $M < p \Rightarrow$ dimension deduction.
- $\sum_{m=1}^{M}\theta_m z_{im} = \sum_{m=1}^{M}\theta_m\sum_{j=1}^{p}\phi_{mj}x_j = \sum_{j=1}^{p}\sum_{m=1}^{M}\theta_m\phi_{mj}x_j = \sum_{j=1}^{p}\beta_j x_j$ where $\beta_j = \sum_{m=1}^{M}\theta_m\phi_{mj}$.
  - ⇒ it's a special case of OLS but with constrains $\beta_j = \sum_{m=1}^{M}\theta_m\phi_{mj}$.

How to choose linear combinations of predictors:

3.1 Principle Components Regression (PCR)

method: apply Principle Components Analysis (PCA) to define linear combinations of the predictors

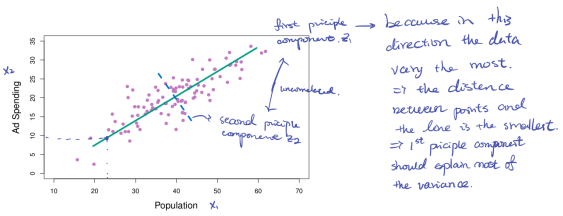
1st principle component: normalized linear combination of predictors with the largest variance.

2nd ⋯⋯⋯: largest variance, subject to being uncorrelated with the 1st

⋮

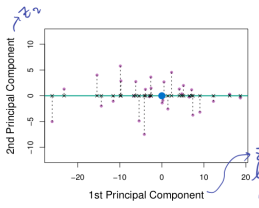
M^{th} principle component: ⋯⋯⋯⋯⋯⋯⋯⋯⋯⋯⋯⋯⋯⋯⋯⋯⋯⋯⋯⋯

⇒ replace the original correlated predictors with a small set of independent principle components that capture their joint variation

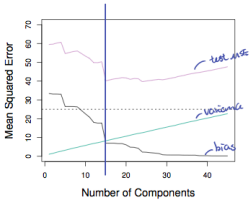


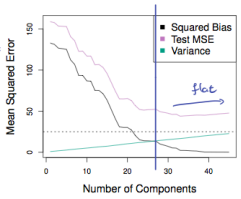
because in this direction the data vary the most.
⇒ the distance between points and the line is the smallest.
⇒ 1st picipie component should explain most of the variance.

tilted →



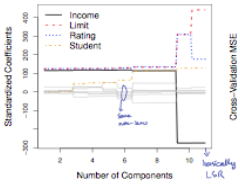
$Z_1 = \phi_{11}X_1 + \phi_{12}X_2$ linear combination of pop. & ad spending

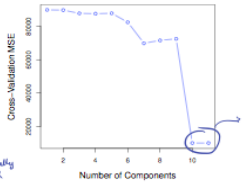




flat

M=15 is chosen →





PCA is not performing any better than LSR for this data set

problem: these directions (or PCs) are identified in an unsupervised way

⇒ there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

⇒ PLS identifies these new features in a supervised way, i.e. it makes use of the response in order to identify new features that are related to the response. ⇒ PLS explains both response and the predictors.

3.2 Partial Least Square (PLS).

(1) standardise p predictors, i.e. $\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\sum_i (x_{ij} - \bar{x}_j)^2}}$

(2) z_1 : set ϕ_{1j} equal to coefficients from simple linear regression of Y onto X_j. one-by-one.

⇒ $z_1 = \sum_j \phi_{1j} X_j$, so ϕ_{1j} is proportional to the correlation between Y and X_j.

⇒ PLS places the highest weight on the predictor that are most related to Y.

(3) take the residuals and repeat the above