Chapter 5 - Resampling methods

1. Cross-validation

Split into training - validation - test set

leave one out: (LOOCV) → fit model n times using $n-1$ datapoints

- $CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$ w/ $MSE_i = (y_i - \hat{y}_{i})^2$ and model is fit w/ the $n-1$ samples leaving the ith one out

- no randomness, always gives the same result (unlike w/ validation set, which is different every if you pick a different validation set)
- uses the whole (almost whole) dataset to fit (unlike if we have a validation set)
- could be expensive to fit

standard for regression:

$$CV_{(n)} = \frac{1}{n}\sum\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2$$

K-fold Cross-validation

- K groups of the dataset, train on $k-1$, validate on 1

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

- less expensive to fit than LOOCV
- some variability depending how we chose the k groups

- LOOCV uses more data to fit than k-fold CV, so is preferred from a bias-perspective
- But LOOCV has a higher variance than k-fold CV
  it is the avg of n very correlated models, while k-fold has the avg of ten correlated models.
  averages of very correlated things have a higher variance

Best is to choose $k=5$ or 10. for k-fold CV

→ For classification problems, just take # misclassifications instead of MSE

II Bootstrap

Generate new datasets w/ repeated sampling from the original dataset
We estimate the value of a parameter for each dataset
we get the mean and the Standard error of the parameter for the bootstrap datasets to have an idea of the true value.

Exercises

1.

$$\min_\alpha \mathrm{Var}(\alpha X + (1-\alpha)Y)$$
$$= \min_\alpha \alpha^2 \mathrm{Var}(X) + (1-\alpha)^2 \mathrm{Var}(Y) + 2\alpha(1-\alpha)\mathrm{Cov}(X,Y)$$
$$= \min_\alpha \alpha^2(\sigma_{xx}^2 + \sigma_{yy}^2 - 2\sigma_{xy}) + \alpha(2\sigma_{xy} - 2\sigma_{yy}^2) + \sigma_{yy}^2$$

derivate $=0$:

$$0 = 2\alpha(\sigma_{xx}^2 + \sigma_{yy}^2 - 2\sigma_{xy}) + 2\sigma_{xy} - 2\sigma_{yy}^2$$

$$\alpha = \frac{\sigma_{yy}^2 - \sigma_{xy}}{\sigma_{xx}^2 + \sigma_{yy}^2 - 2\sigma_{xy}}$$

2. Bootstrap sample n observation

a) p(first bootstrap observation not the j^{th} sample)
$= 1 - 1/n$

b) p(second bootstrap not the j^{th} sample)
$= 1 - 1/n$

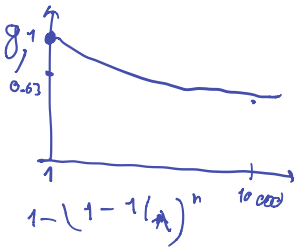
c) p(j^{th} observation is not in the bootstrap)
$= (1-1/n)^n$ (not the j^{th} sample for any samples)

d) p(j^{th} observation is in the sample)
$= 1 - (1-1/n)^n$

e) $n=5$ $\simeq$ 0.67

f) $n=100$ $\simeq$ 0.63

g) $n=10000$ $\simeq$ 0.63



In general, $(1-1/n)^n \longrightarrow \frac{1}{e} \simeq 0.367$

3. a) k-fold cross validation

You divide your data into k subsets of equal size.
For each set, you train on the rest and evaluate the error

on that set

Then you average your error on all your sets to get an estimate.

(b) advantage relative to : validation set
- you get more data to train
- result is less random (depends less on the set)

du
disadvantage:
- more expensive to fit

adv- relative to LOOCV:
- lower variance
- less expensive to fit

disadv:
- less data to train on
- not always the same result

(4) Estimate standard deviation w/ bootstrap:
- Resample from dataset many times to get artificial datasets
- Estimate prediction for each dataset
- Calculate the std dev. of the estimates

6. Linear model selection and regularisation

least squares fit is not the best for linear regression

- if $p \approx n$ or $p > n$ (lots of features)
  - as can result in overfitting
  - or cannot even be used
- if some variables are irrelevant – least squares won't put coeff to 0

alternatives:
- subset selection
- regularisation
- dimension reduction

I Subset selection

Best subset selection

1. M_0 – no predictors, just sample mean.
2. for $k = 1, \dots p$
   (a) fit all $\binom{p}{k}$ models w/ k predictors
   (b) choose best model M_k for all k (smallest RSS or highest R^2)
3. Select best model among $M_0, \dots M_p$ using cross-validated prediction error, C_p (AIC), BIC or adjusted R^2

Stepwise selection

Forward:
1. M_0 – null model
2. For $k = 0, \dots p-1$
   (a) Consider all $p-k$ models that augment the predictors in M_k w/ one additional predictor

(b) Choose the one w/ the highest R^2 or lowest RSS

3. Select best model from $M_0 \dots M_p$ using cross-validated prediction error, C_p(AIC), BIC or adjusted R^2

$\Rightarrow 1 + \frac{p(p+1)}{2}$ models, not 2^p

- faster
- not guaranteed to find optimal model
- can be used if $n < p$, but only $M_0, \dots M_{n-1}$ can be constructed

Backward:

1. M_p - full model
2. For $k = p, \dots 1$
   (a) Consider all k models that contain the predictors in M_k except one predictor
   (b) Choose the one w/ the highest R^2 or lowest RSS
3. Select best model from $M_0 \dots M_p$ using cross-validated pred. err. C_p(AIC), BIC or adjusted R^2

$\Rightarrow 1 + \frac{p(p+1)}{2}$ models, not 2^p

- faster
- not guaranteed to find optimal
- cannot be used if $n < p$

Hybrid approaches:

adding and removing in each step

Choosing the best model

RSS and R^2 is the lowest for the model w/ most features
this is not a good measure to use

2 approaches: I adjustment taking into account the number of features
II estimating the test error w/ validation

1

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

penalty of $2d\hat{\sigma}$ for having RSS

$\hat{\sigma}^2$ is estimated variance for ε (usually w/ all the predictors)

C_p is unbiased estimate of test RSS
if $\hat{\sigma}^2$ is unbiased estimate for σ^2 (not proven here)

choose best model : the one w/ lowest C_p

b) AIC

For large models fit w/ maximum likelihood

For $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$,
MLE is same as least squares so we can use it.

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

So for least squares, $C_p \sim AIC$

c) BIC

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

$\log n > 2$ for $n > 7$, so heavier penalty for large models

d) adjusted R^2

$$\text{adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \qquad \left(\text{instead of } 1 - \frac{RSS}{TSS}\right)$$

For this one, we want the model w/ largest adjusted R^2

Since $TSS = \sum (y_i - \bar{y})^2$ is fix, this is equivalent to minimizing $\frac{RSS}{n-d-1}$

II Validation and CV

Choose model w/ smallest validation / CV error

\+ : direct estimate for test error
less assumptions about model
works for models w/o explicit degrees of freedom
or when hard to estimate σ^2

\- expensive

One-standard-error rule

Calculate standard error for estimated MSE for all models
Select smallest where estimated test error is within one SE
of the lowest point on the curve

meaning if models are similar, ← select simpler model!

6.2 Shrinkage methods – regularising the coeffs

Ridge regression → L2 penalty

$RSS = \sum (y_i - \beta_0 - \sum \beta_j x_{ij})^2$

Instead of minimising RSS, we want to minimise $RSS + \lambda \sum_{j=1}^{p} \beta_j^2$

$\lambda \geq 0$ is a tuning parameter

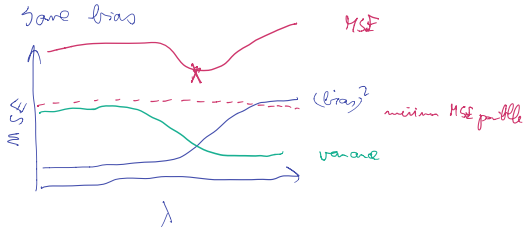
we do not shrink β_0, as this is estimate for the mean!

λ penalising large β values : as λ grows, $\|\hat{\beta}^R_\lambda\|_2$ decreases always
Normal least squares is scale equivariant – multiplying x_j by c
leads to the same model.

THIS IS NOT TRUE FOR RIDGE REGRESSION!!!

STANDARDIZE: $\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum (x_{ij} - \bar{x})^2}}$ (s.t. dev of 1)

Since $E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\varepsilon)$
this helps us reducing the variance little increasing



disadvantage: doesn't set any coeffs to 0 makes it harder to interpret the model always uses all of the predictors

Lasso — tries to overcome the problem of ridge by forcing some params to 0 → variable selection

using (L1 penalty)

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

"sparse models" = not using all variables

Equivalent formulations:

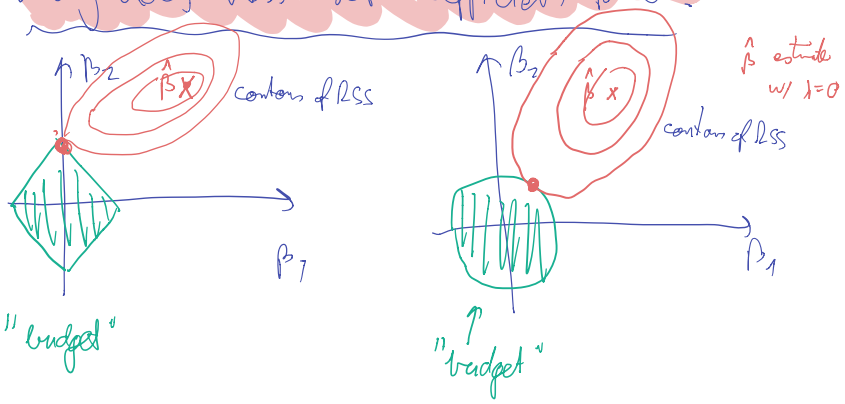
$$\underset{\beta}{\text{minimize}} \; RSS \quad \text{given} \sum \|\beta_j\|_m \leq s \quad = \quad \text{Ridge } (m=2) \quad \text{Lasso } (m=1)$$

$m=1$ for Lasso
$m=2$ for ridge

each λ is equivalent to some value s (meaning they result in same coeffs)

"budget for $\|\beta\|_m$" — as long as least squares is within the budget, it will result in that.

Why does Lasso set coefficients to 0?



because of the sharp edge, Lasso gets an intersection w/ lowest RSS at the contour more often

Ridge vs Lasso

Ridge is better when all variables related

Lasso is better when only few are actually related

⇒ can't know in advance, try both

Bayesian interpretation

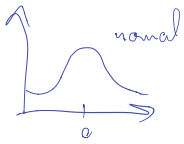
β has a prior $p(\beta)$

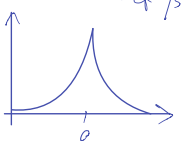
$p(\beta|X,y) \propto f(y|X,\beta)\, p(\beta|X) = f(y|X,\beta)\, p(\beta)$

Assume $y = \beta_0 + X_1\beta_1 + \dots + X_p\beta_p + \varepsilon \quad \varepsilon \sim N(0,\sigma^2)$

and $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$ for some density fn g

- If $g \sim \text{Normal}(0, \text{RSS}(\lambda))$ ⇒ posterior mode of β is given by ridge regression
  [and also the posterior mean]
- If $g \sim \text{Laplace}(0, \text{RSS}(\lambda))$ ⇒ posterior mode of β is given by lasso





⇒ more likely to set to 0 as distribution is peaked there

Selecting the Tuning Parameter λ

Choose a grid for λ values (or s)

using cross-validation, compute error w/ fitted error, then refit model using all values

6.3 - Dimension Reduction Methods

$Z_1, \dots, Z_M$ are linear combinations of $X_1, \dots X_p$ w/ $M \leq p$

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \quad \text{for some } \phi_{1m}, \phi_{pm}, m = 1, \dots M$$

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

using least squares for $\theta_0, \dots \theta_M$

reduces $p+1$ coeff $\beta_0, \dots \beta_p$ to having to fit $M+1 < p+1$ coeffs: $\theta_0, \dots \theta_M$

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\Rightarrow \beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm} \qquad (*)$$

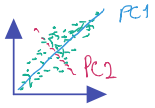
So this is a special case of linear regression where coeff have to obey $(*)$. This is a constraint, so increases bias and decreases variance if $M \ll p$. If $M = p$, and all Z_m are linearly indep, this is not a constraint.

How to get Z_i ? (or ϕ_{jm})

Principal component regression (PCR)

principal components are chosen s.t.

- they are a linear combination of the data
  w/ $\sum_{j=1}^{p} \phi_{jm}^2 = 1$
- s.t. they minimize the variance given that they are uncorrelated / perpendicular to the previous components



We assume that directions in which $X_1, \dots X_p$ vary the most are directly associated w/ Y.

Fitting least squares to $Z_1 \dots Z_M$ is better, because we have most of the signal, and we overfit less due to fewer params

→ NOT A FEATURE SELECTION METHOD, IT USES ALL

closely related to ridge regression

- STANDARDIZE ~~BEFORE~~ PCR
  (when in same units)
- choose number of components w/ cross-validation

Partial Least squares (PLS)

- identify $Z_1, \dots Z_M$, fit w/ least squares

Idea: choose $Z_1, \dots Z_M$ in a supervised way

using Y: those that are related to the response

directions explaining both response and predictors

1) Standardize predictors

2) Compute Z_1 by setting ϕ_{j1} to the coeff from a simple linear regression of Y onto X_j

→ this places highest weight on variables that are most correlated to the response

3) We regress each variable on Z_1, and take residuals (leaving info that is not explained by Z_1)

4) We get $Z_2, \dots Z_M$ as Z_1, using the residuals

Complicated, mostly not better than PCR

High-dimension data

most methods have been developed for $n \gg p$,
having large p is relatively new
→ high-dimensional data

least squares goes wrong if $p \geq n$, because we have many sets of coefs that fit the data perfectly (like line through 1 point)
→ overfit

- C_p, AIC and BIC don't help in high dimensional setting (estimating $\hat{\sigma}^2$ is problematic)
- adjusted R^2 also not, as can easily get adjusted $R^2 = 1$

adding noise features makes the model worse
→ (overfitting)

- if we identify a set of variables that predict the data well, it's not me it's the only set of variables that do so
→ (multicollinearity)
- if $p > n$, don't use SSE, p-values, R^2
(they could be arbitrarily small due to overfitting)

→ use test set error / CV error
MSE / R^2 on independent test set is valid,
but not on training set

Exercises

1) Best subset Forward stepwise Backward stepwise

we get p+1 models for 0, ... p predictors

for k predictors:

smallest training RSS:

X

smallest test RSS:

not sure. least subset selection has the most flexibility, but other methods might be lucky ...

(i) True

(ii) True

(iii) False

(iv) False

(v) False

2) Lasso relative to least squares: (iii) Ridge relative to least squares (iii) non-linear relative to L.S. (ii)

3) training RSS: iv

4) iii

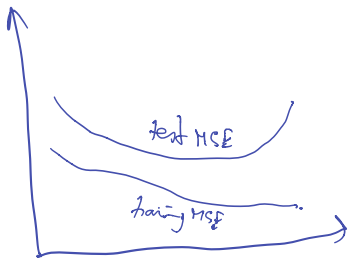
ii

test MSE : ii

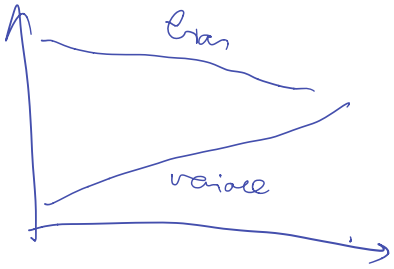
Variance : iii iv

Bias : iv

Irreducible error : v v

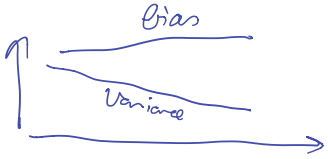


3



2

1



4

5) $n=2$ $p=2$

x_{ij} : $i: 1 \to n$, $j: 1 \to p$

$x_{11} = x_{12} = x_1$ $\quad y_1 + y_2 = 0$

$x_{21} = x_{22} = x_2$ $\quad x_{11} + x_{21} = 0$

$x_{12} + x_{22} = 0$

$\hat{\beta}_0 = 0$

a) Ridge :

$$\min \left(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12}\right)^2 + \lambda\left(\hat{\beta}_1^2 + \hat{\beta}_2^2\right) + \left(y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22}\right)^2$$

$$\Rightarrow \min \left(y_1 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_1\right)^2 + \left(y_2 - \hat{\beta}_1 x_2 - \hat{\beta}_2 x_2\right)^2 + \lambda\left(\hat{\beta}_1^2 + \hat{\beta}_2^2\right)$$

derivative:

$$\frac{\partial}{\partial \beta_1} : -2(y_1 - \hat\beta_1 x_1 - \hat\beta_2 x_1)\, x_1$$
$$-2(y_2 - \hat\beta_1 x_2 - \hat\beta_2 x_2)\, x_2$$
$$+ 2\lambda \hat\beta_1$$

$$\lambda \hat\beta_1 = y_1 x_1 + y_2 x_2 - \hat\beta_1 x_1^2 - \hat\beta_2 x_1^2 - \hat\beta_1 x_2^2 - \hat\beta_2 x_2^2$$

(I) $\hat\beta_1 (\lambda + x_2^2 + x_1^2) = y_1 x_1 + y_2 x_2 - \hat\beta_2 (x_1^2 + x_2^2)$

(II) $\hat\beta_2 (\lambda + x_2^2 + x_1^2) = y_1 x_1 + y_2 x_2 - \hat\beta_1 (x_1^2 + x_2^2)$

I − II $\lambda (\hat\beta_1 - \hat\beta_2) = 0 \Rightarrow \hat\beta_1 = \hat\beta_2$

c) Lasso:

$$\min. \; (y_1 - \hat\beta_1 x_1 - \hat\beta_2 x_1)^2 + (y_2 - \hat\beta_1 x_2 - \hat\beta_2 x_2)^2 + \lambda (|\hat\beta_1| + |\hat\beta_2|)$$

or $\min. \; (y_1 - \hat\beta_1 x_1 - \hat\beta_2 x_1)^2 + (y_2 - \hat\beta_1 x_2 - \hat\beta_2 x_2)^2$ given $|\hat\beta_1| + |\hat\beta_2| \le s$

nice $x_1 = -x_2$, this is the same as
$y_1 = -y_2$

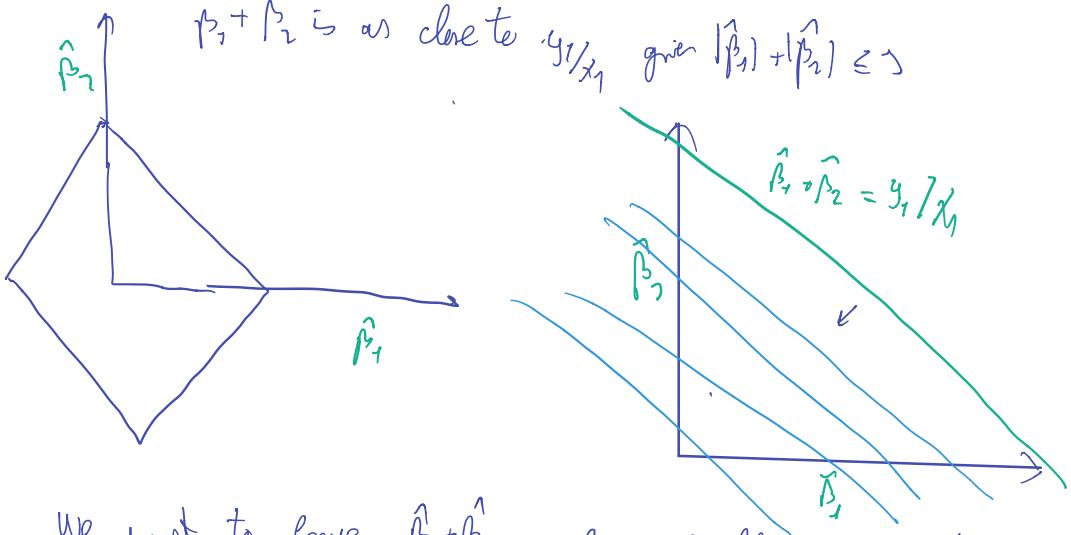
$$\min. \; [y_1 - (\hat\beta_1 + \hat\beta_2) x_1]^2 \quad \text{given } |\hat\beta_1| + |\hat\beta_2| \le s$$

So $\min \; |y_1 - (\hat\beta_1 + \hat\beta_2) x_1|$

this is true when $\hat{\beta}_1 + \hat{\beta}_2 = \frac{y_1}{x_1}$

So we optimize when

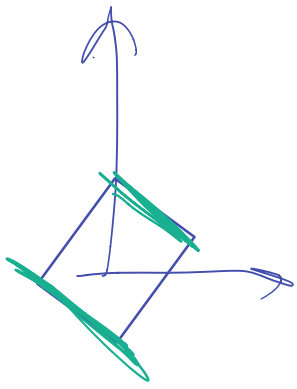
$\hat{\beta}_1 + \hat{\beta}_2$ is as close to y_1/x_1 given $|\hat{\beta}_1| + |\hat{\beta}_2| \leq s$



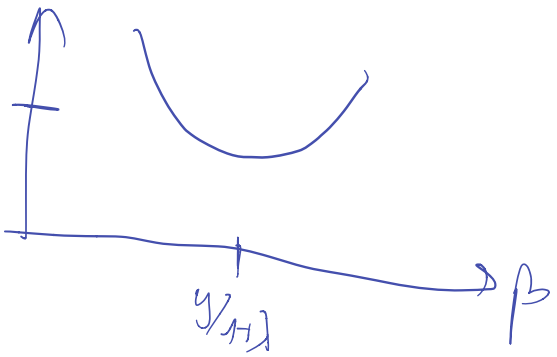
We want to have $\hat{\beta}_1 + \hat{\beta}_2$ as big as possible but to intersect the diamond, so the solution is the first intersection w/ the diamond.

If $\frac{y_1}{x_1} > 0$, top right edge,

if $\frac{y_1}{y_1} < 0$, left bottom edge



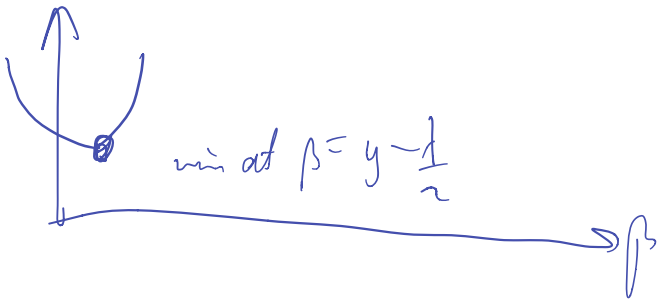
(a) $(y-\beta)^2 + \lambda\beta^2$ $\quad p=1$ $n=1$ ridge



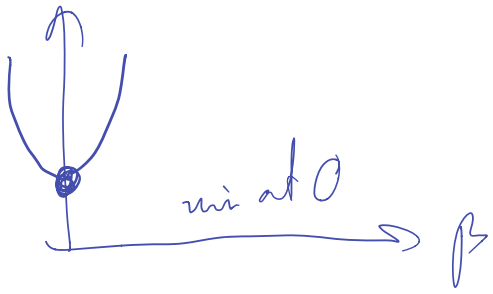
$\lambda > 0$

$\beta^2(\lambda+1) - 2\beta y + y^2$

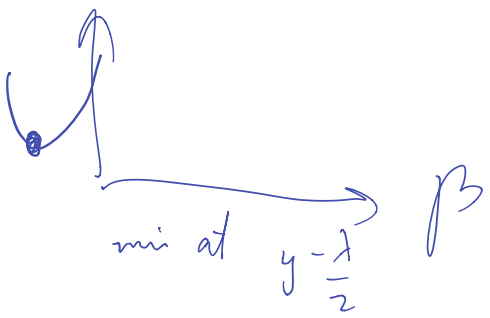
(b) $(y-\beta)^2 + \lambda|\beta|$



$y > \lambda/2$



$|y| < \lambda/2$



$y < -\lambda/2$

min at $-2(y-\beta) + \lambda s \quad (s = -1 \text{ or } +1)$

assume $\beta > 0 \quad \lambda - 2y + 2\beta = 0$

min at $\beta = y - \frac{\lambda}{2}$.

so min at $\beta = y - \frac{\lambda}{2}$ when $y > \lambda/2$

assume $\beta < 0 \quad -\lambda - 2y + 2\beta = 0$

$\beta = y + \frac{\lambda}{2} \rightarrow y < -\frac{\lambda}{2}$

min at $\beta = y + \frac{\lambda}{2}$ when $y < -\frac{\lambda}{2}$

what if $|y| < \frac{\lambda}{2}$?

$$\begin{cases} \beta^2 - (2y+\lambda)\beta + y^2 & \text{if } \beta \le 0 \\ \beta^2 + (\lambda - 2y)\beta + y^2 & \text{if } \beta > 0 \end{cases}$$

min at $2\beta = 2y+\lambda \to \beta = \frac{y+\lambda}{2}$

But if $|y| < \frac{\lambda}{2}$ and $\lambda > 0$, $\beta > 0$, so

this has ↯ min at 0

same for $\beta > 0$: min at $y - \frac{\lambda}{2} < 0$, so

this has min at 0

$\Rightarrow$ min at 0 for $(y-\beta)^2 + \lambda|\beta|$

↯ $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \qquad \varepsilon_i \sim N(0,\sigma^2)$ iid

likelihood:

$$p(y_i \mid x_{ij}, \beta_j) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - (\beta_0 + \sum_j x_{ij}\beta_j))^2}{2\sigma^2}}$$

b) Assume $\beta_1, \dots \beta_n$ are iid Laplace$(0,\theta)$

$$p(\beta) = \frac{1}{2b} \exp -\frac{|\beta|}{b}$$

posterior for β:

$$p(\beta \mid y_i) = \frac{p(y|\beta)\, p(\beta)}{p(y)} \propto p(y|\beta)\, p(\beta)$$

$$\propto = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{2b}\right) \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})\right]^2 - \frac{|\beta|}{b}\right)$$

So mode for β is maximizing the posterior

take logs and drop irrelevant terms:

$$\max \quad -\frac{1}{2\sigma^2} \Sigma \left[y_i - (\beta_0 + \Sigma \beta_j x_{ij})\right]^2 - \frac{|\beta|}{b}$$

$$= \min \quad \Sigma \left[y_i - (\beta_0 + \Sigma \beta_j x_{ij})\right]^2 + \frac{\Sigma |\beta_j|}{b} \cdot 2\sigma^2$$

$$= \min \quad RSS + \frac{2\sigma^2}{b} \Sigma |\beta|$$

take $\lambda = \frac{2\sigma^2}{b}$, this is the lasso solution.

d) $\beta_1 \dots \beta_p \sim N(0, c)$

posteria de β:

–

The posterior distributed according to Normal distribution with mean 0 and variance c is:

$$f(\beta \mid X, Y) \propto f(Y \mid X, \beta) p(\beta \mid X) = f(Y \mid X, \beta) p(\beta)$$

Our probability distribution function then becomes:

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

Substituting our values from (a) and our density function gives us:

$$\begin{aligned} f(Y \mid X, \beta) p(\beta) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2\right)\left(\frac{1}{\sqrt{2c\pi}}\right)^p \exp\left(-\frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{2c\pi}}\right)^p \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 - \frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right) \end{aligned}$$

e

Like from part c, showing that the Ridge Regression estimate for β is the mode and mean under this posterior distribution is the same thing as showing that the most likely value for β is given by the lasso solution with a certain λ.

We can do this by taking our likelihood and posterior and showing that it can be reduced to the canonical Ridge Regression Equation 6.5 from the book.

Let's start by simplifying it by taking the logarithm of both sides:

Once again, we can take the logarithm of both sides to simplify it:

$$\begin{aligned} \log f(Y \mid X, \beta) p(\beta) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{2c\pi}}\right)^p \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 - \frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right) \\ &= \log\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{2c\pi}}\right)^p\right] - \left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right) \end{aligned}$$

We want to maximize the posterior, this means:

$$\arg\max_{\beta} f(\beta \mid X, Y) = \arg\max_{\beta} \log\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{2c\pi}}\right)^p\right] - \left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right)$$

Since we are taking the difference of two values, the maximum of this value is the equivalent to taking the difference of the second value in terms of β. This results in:

$$\begin{aligned} &= \arg\min_{\beta}\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right) \\ &= \arg\min_{\beta}\left(\frac{1}{2\sigma^2}\right)\left(\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \frac{\sigma^2}{c}\sum_{i=1}^{p}\beta_i^2\right) \end{aligned}$$

By letting $\lambda = \sigma^2/c$, we end up with:

$$\begin{aligned} &= \arg\min_{\beta}\left(\frac{1}{2\sigma^2}\right)\left(\sum_{i=1}^{n}\left[Y_i - (\beta_0 + \sum_{j=1}^{p}\beta_j X_{ij})\right]^2 + \lambda\sum_{i=1}^{p}\beta_i^2\right) \\ &= \arg\min_{\beta} \text{RSS} + \lambda\sum_{i=1}^{p}\beta_i^2 \end{aligned}$$

Since we find our posteria is normal, mean equals median