



05. Ridge, Lasso and ElasticNet



Jonathan Hersh (Chapman University Argyros School of Business)

Outline

- 1. Ridge Regression**
- 2. Cross-Validation**
- 3. Lasso regression**
- 4. ElasticNet!**

Occam's Razor and Statistics

- In statistics Parsimony is the analogue of Occam's razor
- Between two equal models (performance-wise) we prefer the simpler one
- Simpler = fewer variables
- But which variables are superfluous?! (unnecessary)
- Lasso and Ridge are key methods to derive parsimonious linear models

CORE PRINCIPLES IN RESEARCH



OCCAM'S RAZOR

"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."

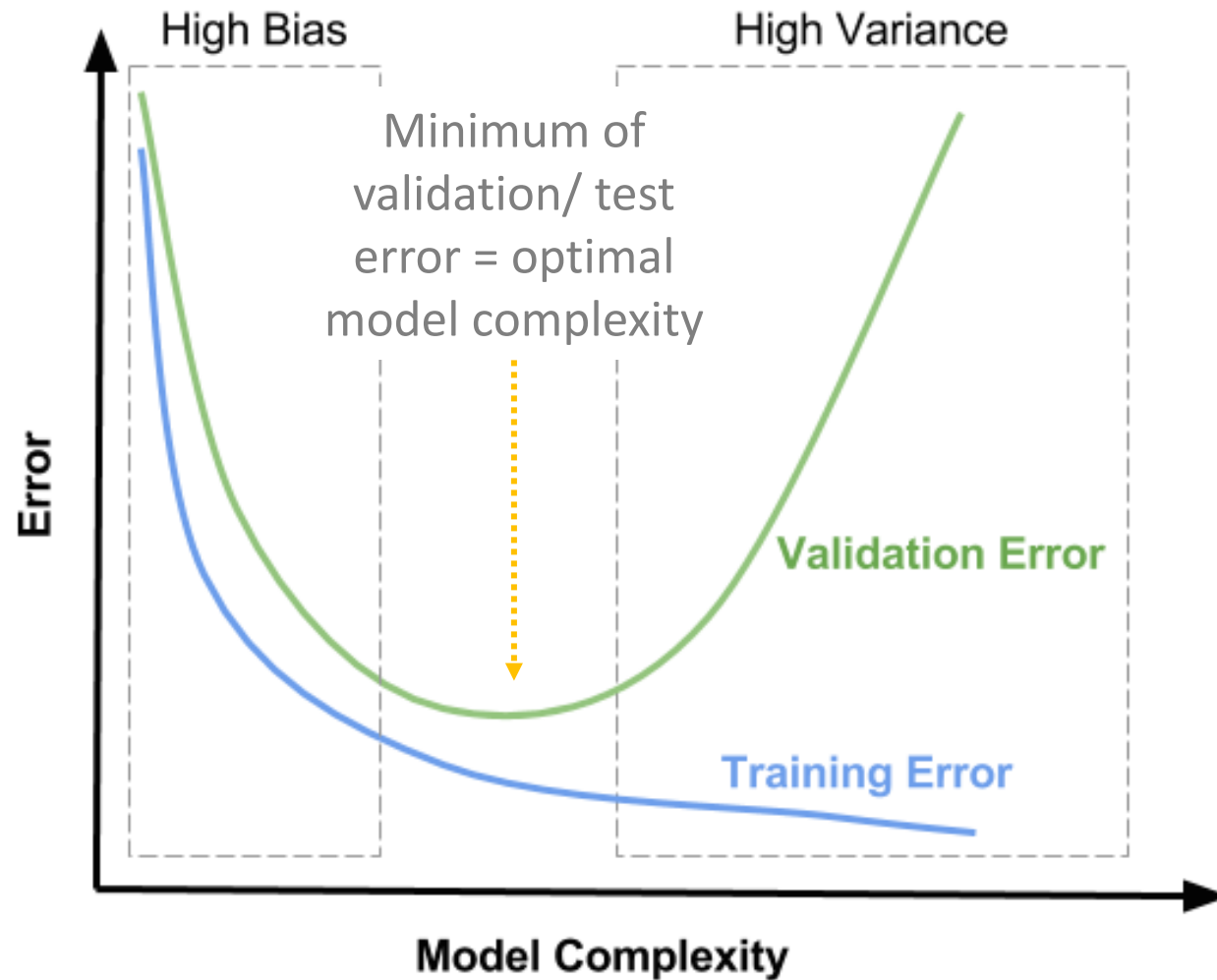


OCCAM'S PROFESSOR

"WHEN FACED WITH TWO POSSIBLE WAYS OF DOING SOMETHING, THE MORE COMPLICATED ONE IS THE ONE YOUR PROFESSOR WILL MOST LIKELY ASK YOU TO DO."

WWW.PHDCOMICS.COM

Recall Bias-Variance Tradeoff




- More model complexity test performance, but beyond a certain point it can increase test/validation error
- Note that training error always increases with model complexity!
- Key is determining optimal model complexity (in linear models, more complexity = more variables)

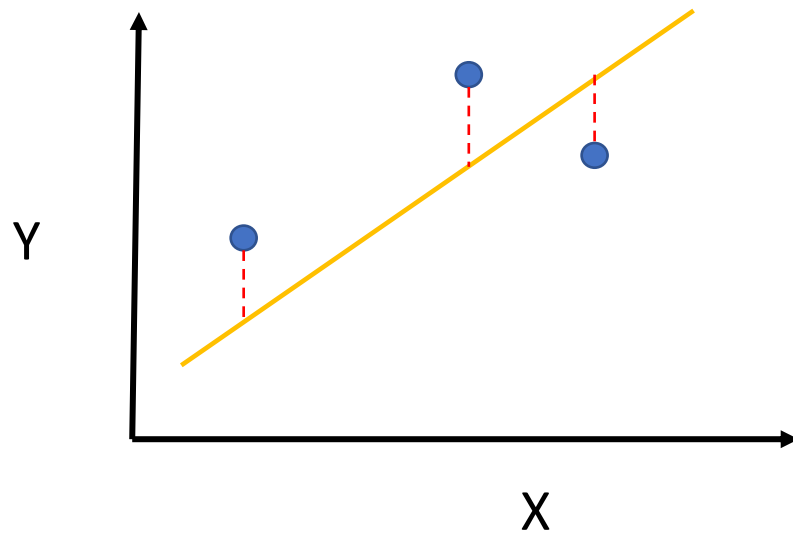
Ridge Regression



Least Squares (OLS Estimator)

$$\hat{\beta} \text{ minimizes: } \sum_{i=1}^N (y_i - \beta_0 - x_{i1}\beta_1)^2$$


Least squares minimizes the sum of squared residuals (e.g. $y_i - \hat{y}$)

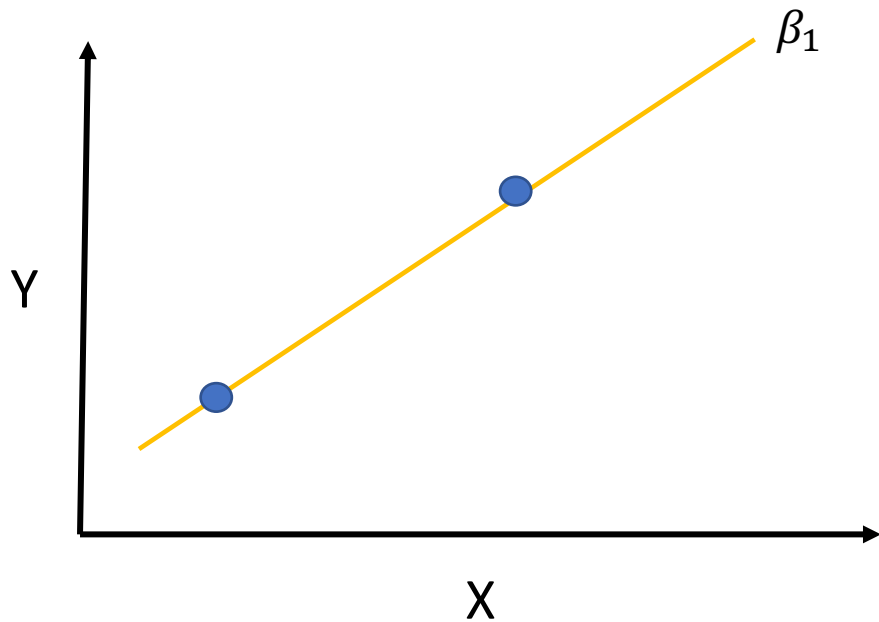


Let's set $K = 1$ (i.e. one explanatory variable to make this easier)

Visually, the slope (β_1) minimizes the difference between the points and the yellow line (red lines)

Ridge Regression Idea

$\hat{\beta}_{ridge}$ minimizes: *residuals* + $\lambda \cdot (\beta_1)^2$

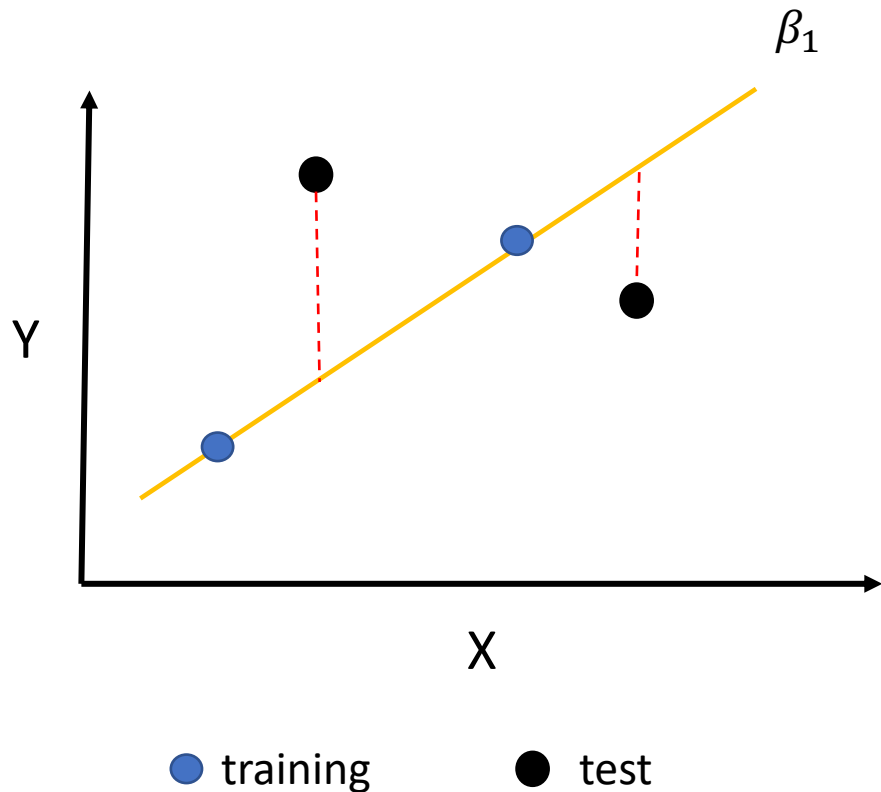


Ridge regression in contrast minimizes the OLS residuals plus the squared slope of beta times λ

Why is this smart? Let's take the example where we only have two data points.

Our best fit line describes the points perfectly and our residuals = 0. Our bias is zero!

OLS Bias and Variance



Bias = 0, which is good

What about our variance?

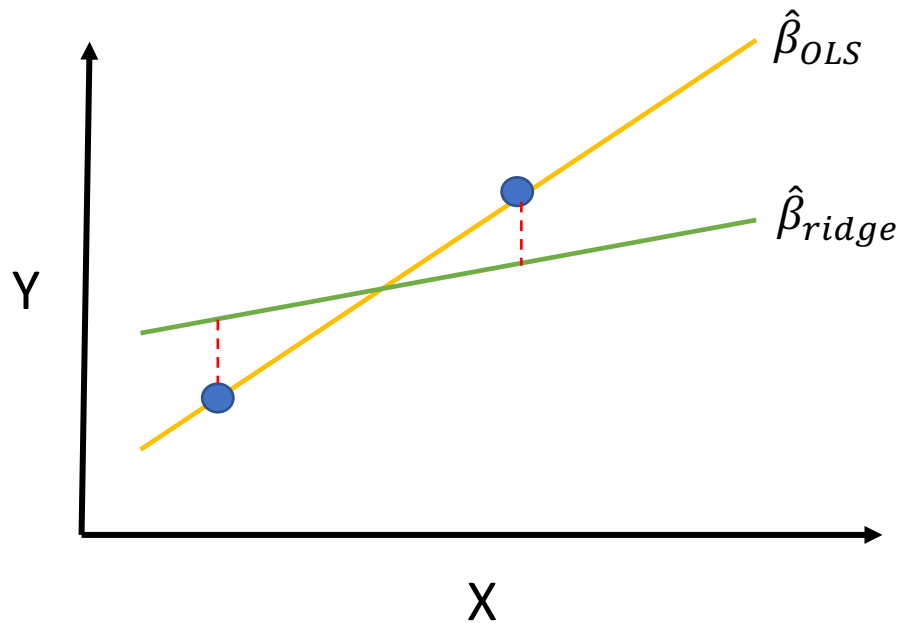
Imaging we had the following data in our test set

Then our test error would be the lines in red, and our variance would be high

Intuition: increasing bias can often reduce variance

Ridge Regression Idea

$\hat{\beta}_{ridge}$ minimizes: $residuals + \lambda \cdot (slope)^2$



Now let $\lambda = 1$. $\hat{\beta}_{ridge}$ minimizes:

$$\begin{aligned} & residuals + 1 \cdot (\beta_1)^2 \\ &= residuals + 1 \cdot (\beta_1)^2 \end{aligned}$$

Suppose the OLS slope = 2

then $\hat{\beta}_{ridge}$ would choose to have some positive residuals to because

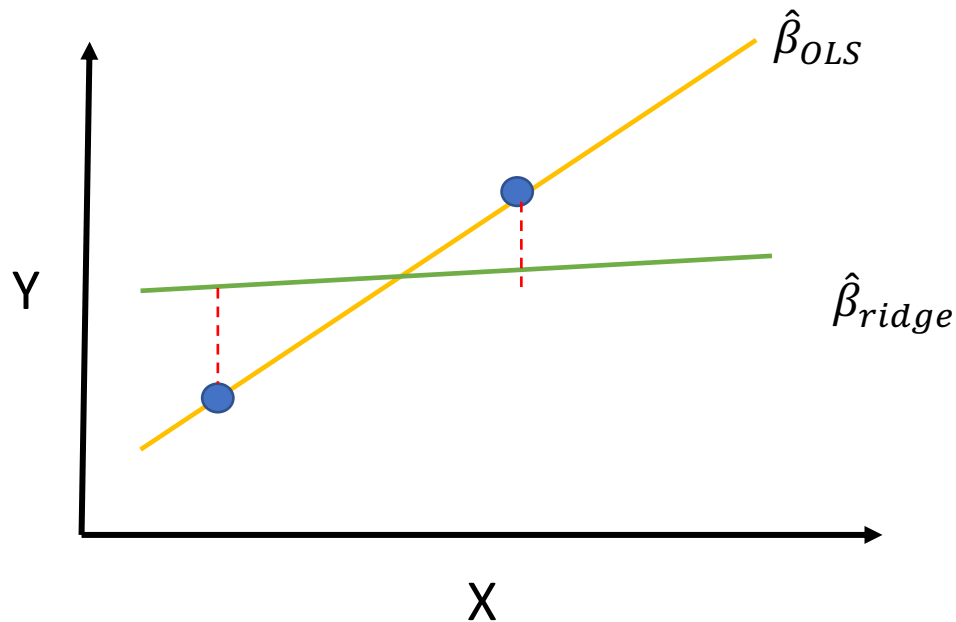
$$\text{residuals_ridge} + 1 \cdot (1)^2$$

is likely less than $0 + (2)^2 = 4$ (residuals with OLS plus lambda penalty)

Aka: we accept a little bias (higher residuals) for less variance (better test performance)

Ridge Regression and Lambda

$\hat{\beta}_{ridge}$ minimizes: $residuals + \lambda \cdot (slope)^2$



What if we set $\lambda = 100$?

$\hat{\beta}_{ridge}$ minimizes:

$$residuals + 100 \cdot (\beta_1)^2$$

Here ridge will have to accept very high residuals in order to avoid high slope penalty

$\hat{\beta}_{ridge}$ will set slope to a very small amount (ex = 0.001) and we get:

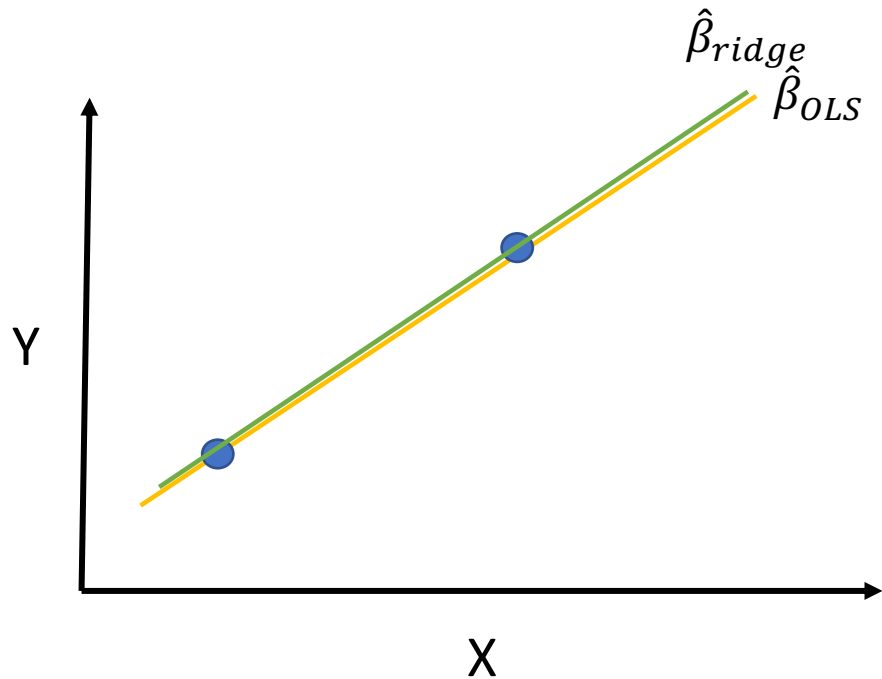
$$residuals + 100 \cdot (0.001)^2$$

$$= residuals + 0.0001$$

E.g. we accept a lot of bias but low variance

Ridge Regression and Lambda

$\hat{\beta}_{ridge}$ minimizes: $residuals + \lambda \cdot (slope)^2$



What if we set $\lambda = 0$?

$\hat{\beta}_{ridge}$ minimizes:

$$residuals + 0 \cdot (\beta_1)^2$$

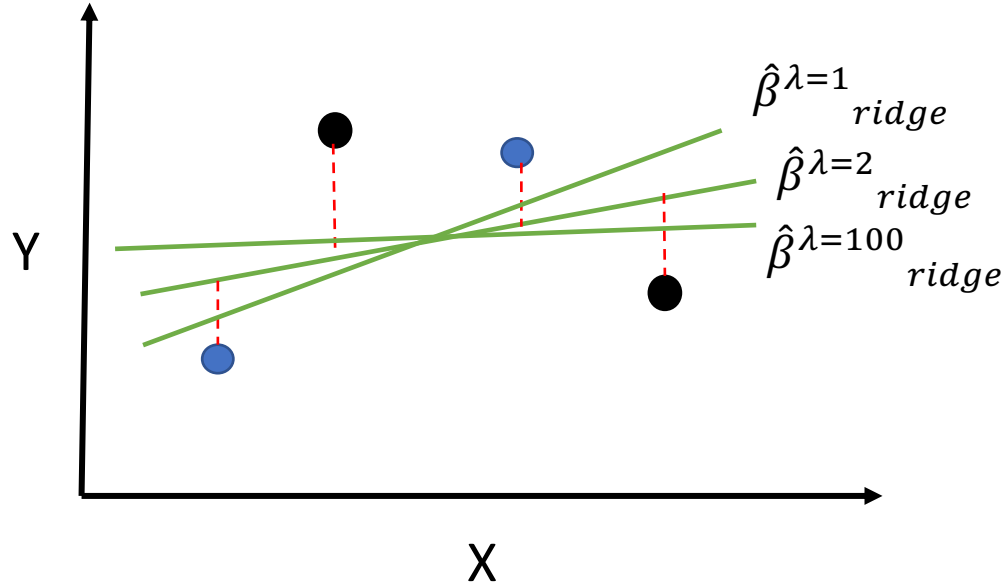
$$= residuals + 0$$

That's just the OLS estimator and $\hat{\beta}_{ridge} = \hat{\beta}_{OLS}$

E.g. lower the penalty on lambda the closer ridge is to OLS.

Larger $\lambda \Rightarrow$ More Penalization, Smaller Coefficients

$\hat{\beta}_{ridge}$ minimizes: *residuals* + $\lambda \cdot (\text{slope})^2$



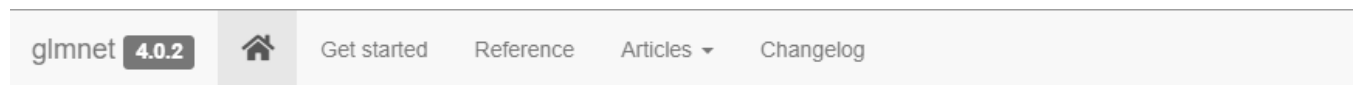
So how do we choose λ ?

In practice we estimate a many models with many different values of λ

We pick a min and max lambda (say 0 and 100), then choose some points in-between

Optimal λ^* minimizes cross-validated error

Ridge Regression in R: glmnet and glmnetUtils



Lasso and Elastic-Net Regularized Generalized Linear Models

We provide extremely efficient procedures for fitting the entire lasso or elastic-net regularization path for linear regression (gaussian), multi-task gaussian, logistic and multinomial regression models (grouped or not), Poisson regression and the Cox model. The algorithm uses cyclical coordinate descent in a path-wise fashion. Details may be found in Friedman, Hastie, and Tibshirani (2010), Simon et al. (2011), Tibshirani et al. (2012), Simon, Friedman, and Hastie (2013).

Version 3.0 is a major release with several new features, including:



Introduction to glmnetUtils

The [glmnetUtils package](#) provides a collection of tools to streamline the process of fitting elastic net models with [glmnet](#). I wrote the package after a couple of projects where I found myself writing the same boilerplate code to convert a data frame into a predictor matrix and a response vector. In addition to providing a formula interface, it also features a function `cva.glmnet` to do crossvalidation for both α and λ , as well as some utility functions.

The formula interface

The interface that glmnetUtils provides is very much the same as for most modelling functions in R. To fit a model, you provide a formula and data frame. You can also provide any arguments that glmnet will accept. Here are some simple examples for different types of data:

```
# least squares regression
(mtcarsMod <- glmnet(mpg ~ cyl + disp + hp, data=mtcars))
```

- glmnet quickly estimates Ridge and Lasso models
- It's one of the best package in R or any language (ported to python in 2015) but can be difficult to work with
- glmnetUtils is a helper package that makes our lives much easier
- Make sure to install both glmnet and glmnetUtils!

Ridge Model with glmnetUtils

```
ridge_mod <- cv.glmnet(any_bank_account ~ urban + tenureTypeOwn + outerWallsPoor  
  + toiletPoor + elecGrid + bedrooms + aircon + fridges  
  + micros + washers + stereos + DVDplayers + TVs +  
  + cellphones + computers + vehicles + cable +  
  + internet + numHHmem + numDep + numChildren,  
  data = LFS_train,  
  weights = Weight,  
  family = "binomial",  
  # note alpha = 0 sets ridge!  
  alpha = 0)
```

```
> print(ridge_mod$lambda.min)
```

```
[1] 0.09385864
```

```
> #
```

```
> print(ridge_mod$lambda.1se)
```

```
[1] 0.797567
```

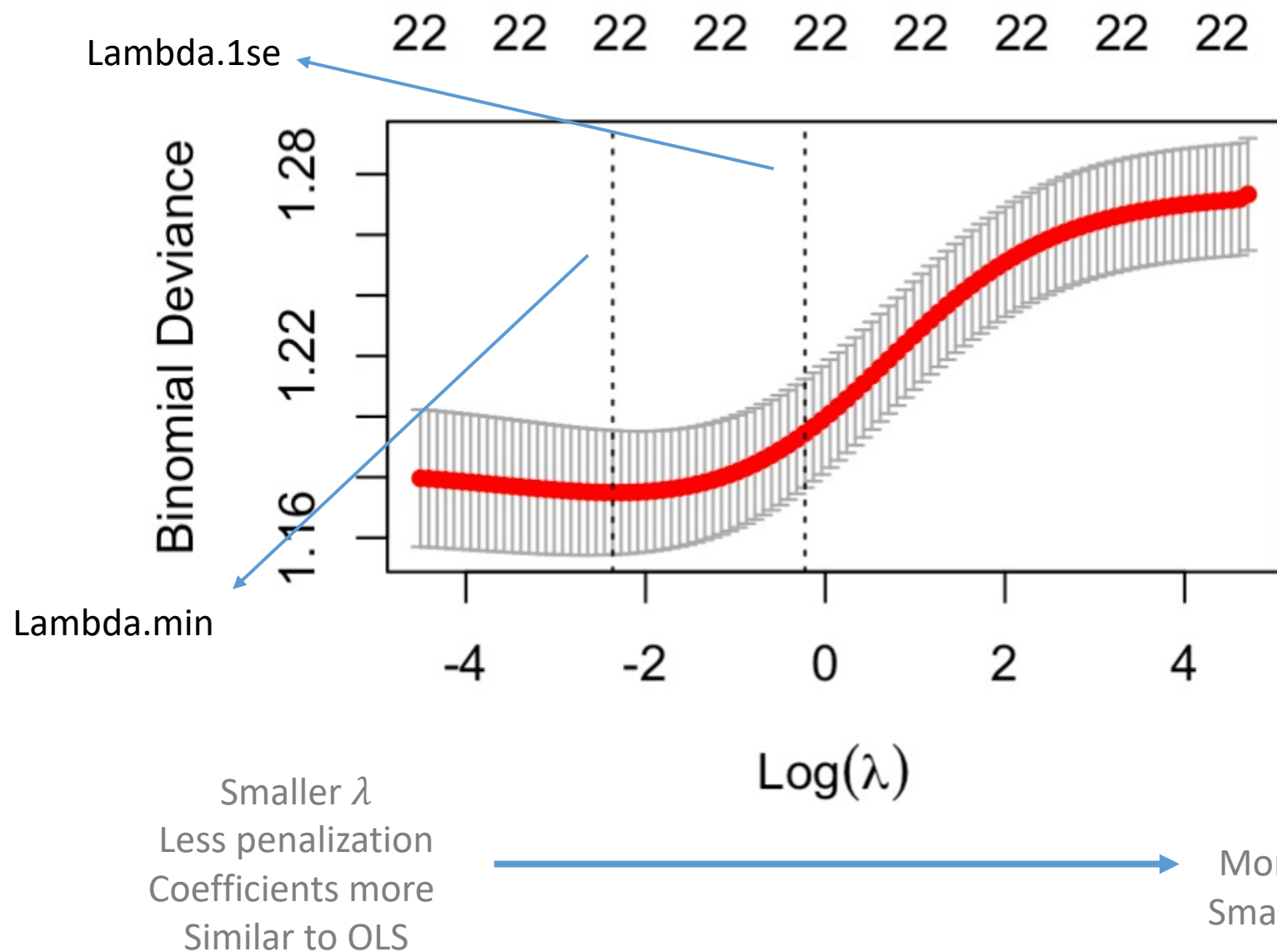
- cv.glmnet estimates a lasso or ridge model. Automatically performs cross-validation to select optimal lambda!
- We must set alpha = 0 to signify ridge model

- lambda.min stores the value of lambda that minimizes cross-validated error

- lambda.1se stores the value of lambda that minimizes cross-validated error plus one estimated standard error

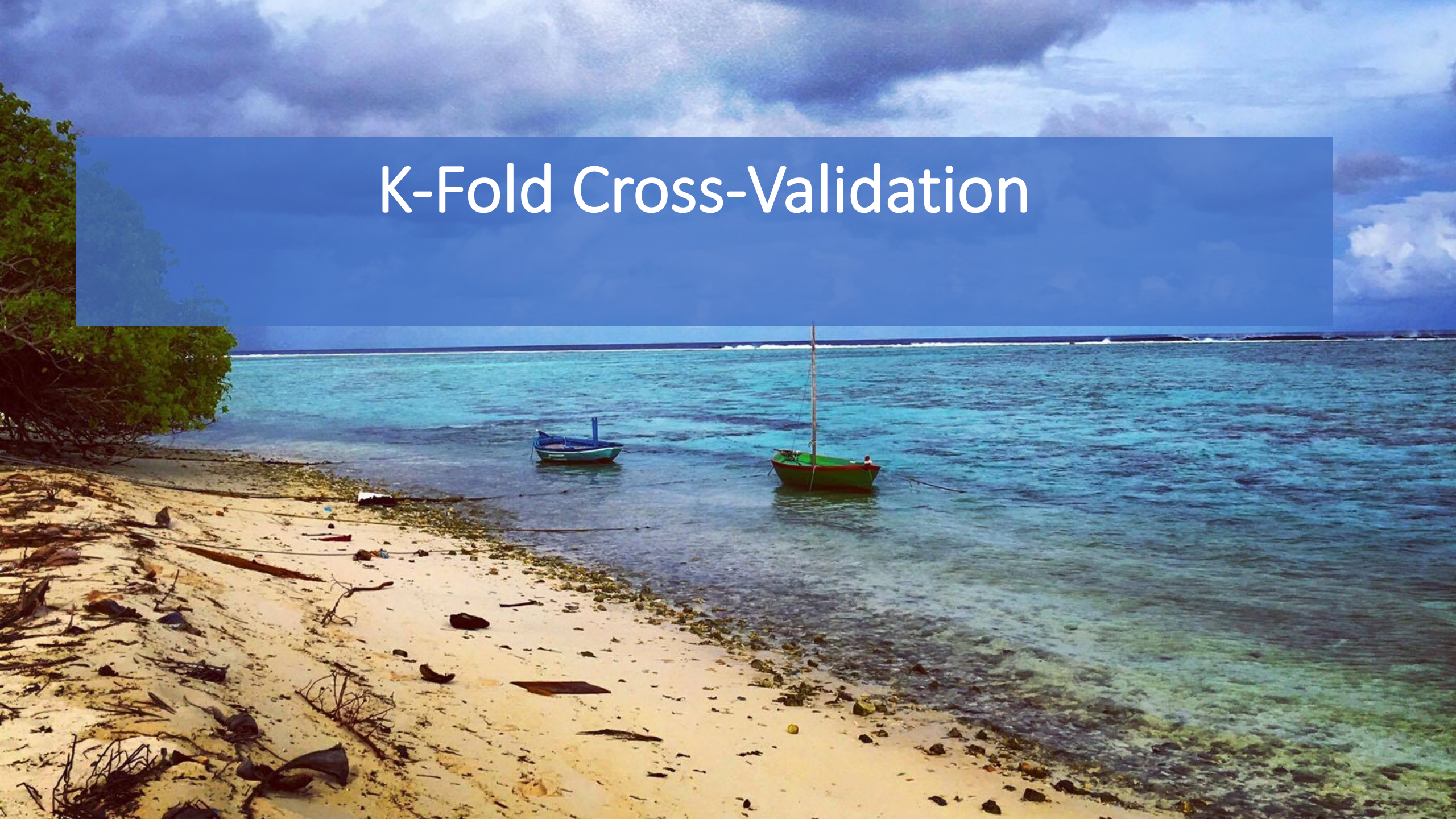
- Why the difference? Lambda.min gives the best performing value, lambda.1se add extra penalization for more parsimony

Cross-Validated MSE Plot As A Function of Lambda



- `plot(model_object)` calls the MSE plot
- This shows how the cross-validated MSE (y-axis) varies as we increase lambda (penalization)
- Model defaults to lambda.1se but either can be appropriate

K-Fold Cross-Validation

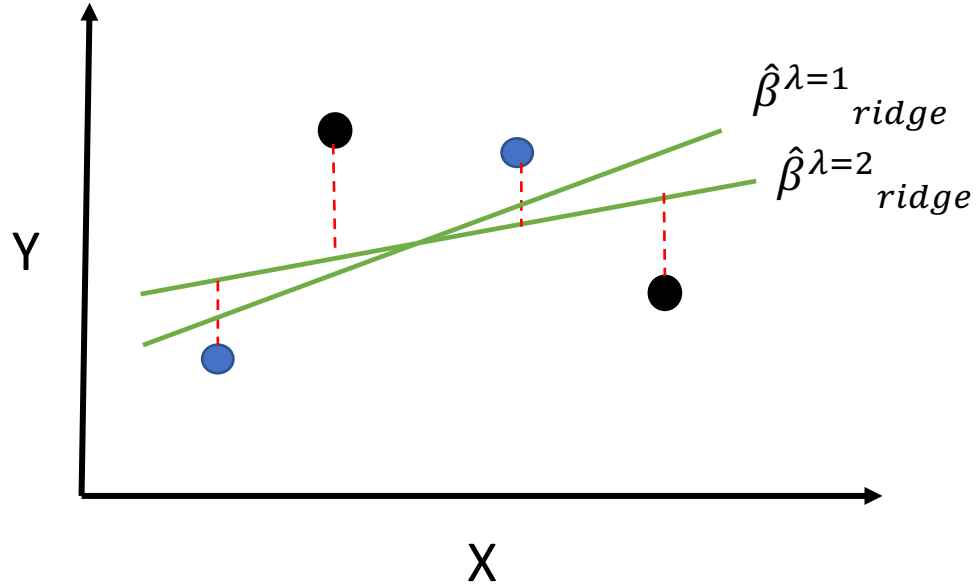


How to choose λ ?

- λ determines how parsimonious our final model is
 - Higher λ \rightarrow more parsimonious = fewer variables
 - Lower λ \rightarrow less parsimonious = more variables
- In practice, we will estimate several values of λ and see which best approximates out of sample-fit.
- **Cross-validation** is a clever technique that approximates out of sample fit.
- Why not just estimate against the test set? We want to save our test set to estimate final “tuned” model performance after we’ve chosen λ

Choosing λ for Ridge?

$\hat{\beta}_{ridge}$ minimizes: *residuals* + $\lambda \cdot (\text{slope})^2$



So how do we choose λ ?

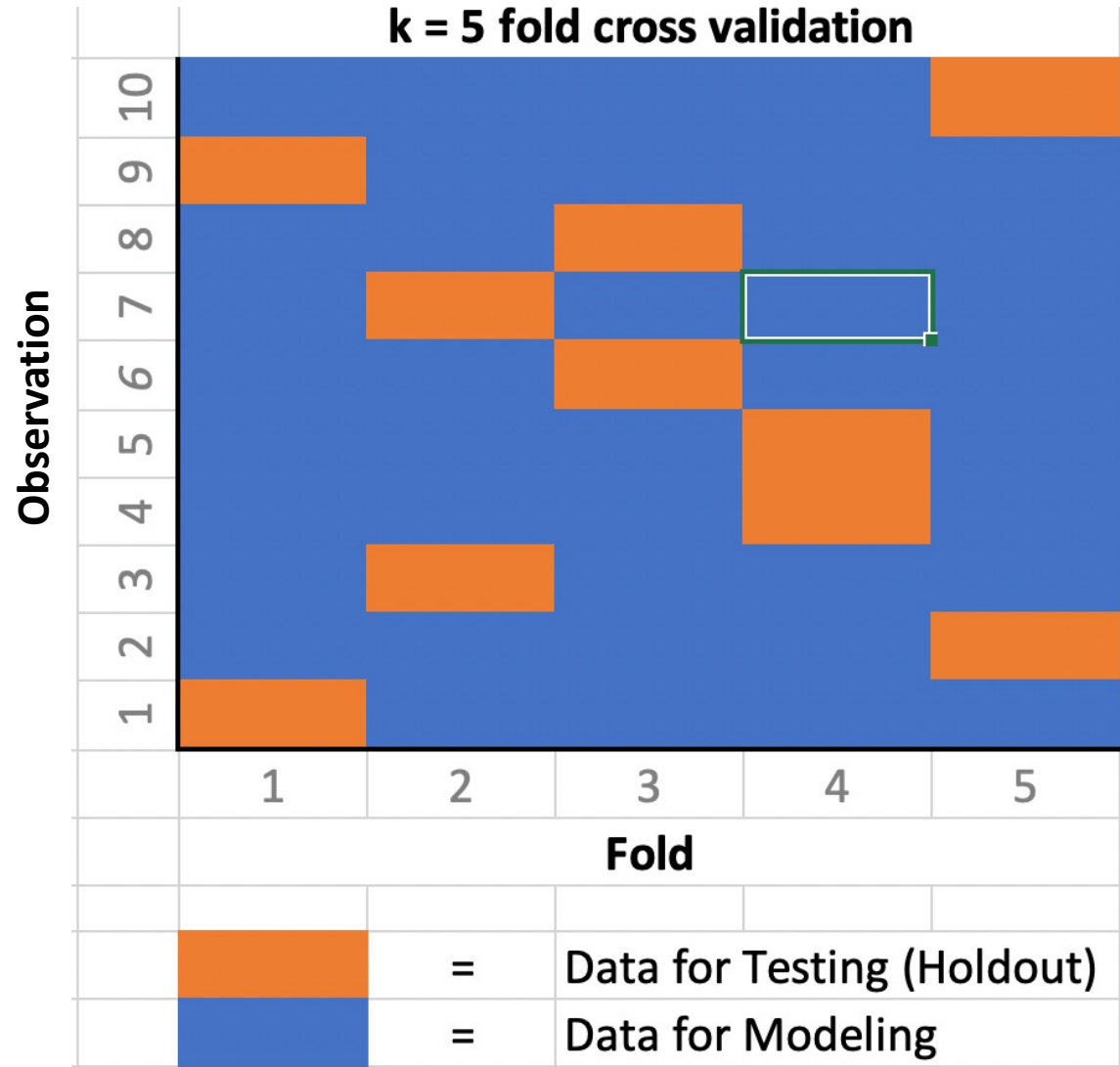
In practice we estimate a many models with many different values of λ

We pick a min and max lambda (say 0 and 10), then choose some points in-between

Our **optimal** λ^* is the lambda that minimizes cross-validated error

K-Fold Cross-Validation

- With k-fold **cross-validation** we first partition (divide) data into K distinct groups
- Fit a model using data excluding group 1, use that model to predict into group 1.
- Fit a model using data excluding group 2, etc
- Proceed until we have y-hats for every group



Resampling: K-Fold Cross-Validation

- We start by randomly assigning each data point to one of k folds
- Here we are setting $k = 3$
- We fit a model excluding data from fold 1
- That model is used to predict into fold 1

\hat{y}^{KCV}	Fold	mpg	cyl	Displ
	3	20	4	3
	2	15	6	5
	3	12	4	2.4
11	1	10	8	4.6
	2	14	6	3
22	1	25	4	2

$X^{-\{1\}}$: X excluding fold 1

Resampling: K-Fold Cross-Validation

- Here we are setting $k = 3$
- Next we fit a model excluding observations in fold 2
- That model is used to predict into fold 2

\hat{y}^{KCV}	Fold	mpg	cyl	Displ
	3	20	4	3
18	2	15	6	5
	3	12	4	2.4
11	1	10	8	4.6
15	2	14	6	3
22	1	25	4	2

$X^{-\{2\}}$: X excluding fold 2

Resampling: K-Fold Cross-Validation

- Here we are setting $k = 3$
- Next we fit a model excluding observations in fold 3
- That model is used to predict into fold 3

$\hat{f}_{X-\{3\}}(X^{-\{3\}})$

$\hat{y}_{\{3\}}^{KCV} = \hat{f}_{X^{\{3\}}}(X^{\{3\}})$

\hat{y}^{KCV}	Fold	mpg	cyl	Displ
22	3	20	4	3
18	2	15	6	5
12	3	12	4	2.4
11	1	10	8	4.6
15	2	14	6	3
22	1	25	4	2

$X^{-\{3\}}$: X excluding fold 3

K-Fold CV Versus LOOCV

- Advantages of K-Fold CV over LOOCV
 - Only need to estimate K models
- Disadvantages
 - Higher variance (more uncertainty in \hat{y}_{hats})
- Because of computational cost K-Fold CV more commonly used

\hat{y}^{KCV}	mpg	cyl	Displ
18	20	4	3
16	15	6	5
12	12	4	2.4
11	10	8	4.6
11	14	6	3
22	25	4	2

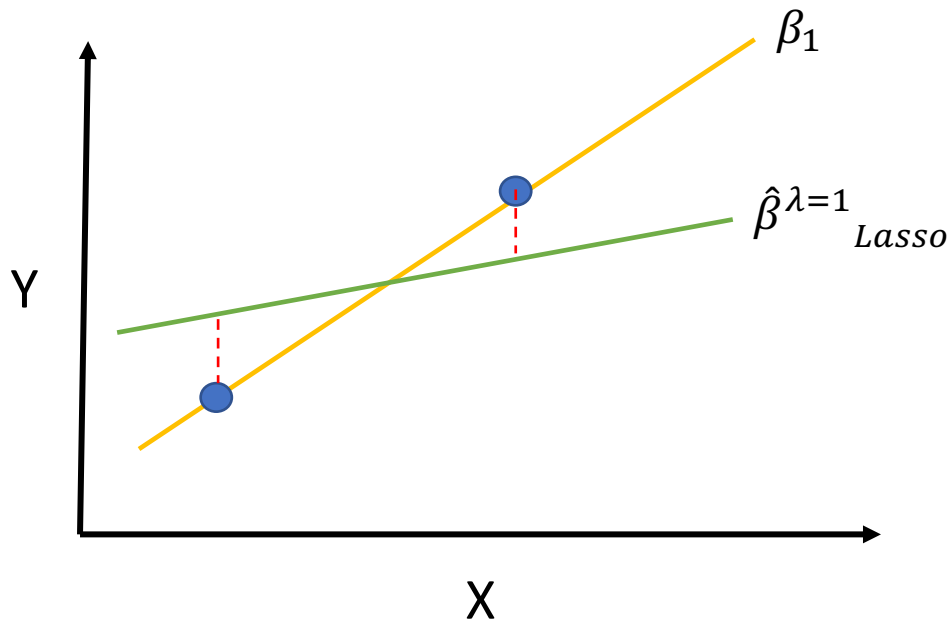
$$MSE_{KCV} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i^{KCV})^2$$

Lasso Regression



Lasso Regression Idea

$\hat{\beta}_{Lasso}$ minimizes: *residuals* + $\lambda \cdot (|\beta_1| + |\beta_2| + \dots + |\beta_k|)$



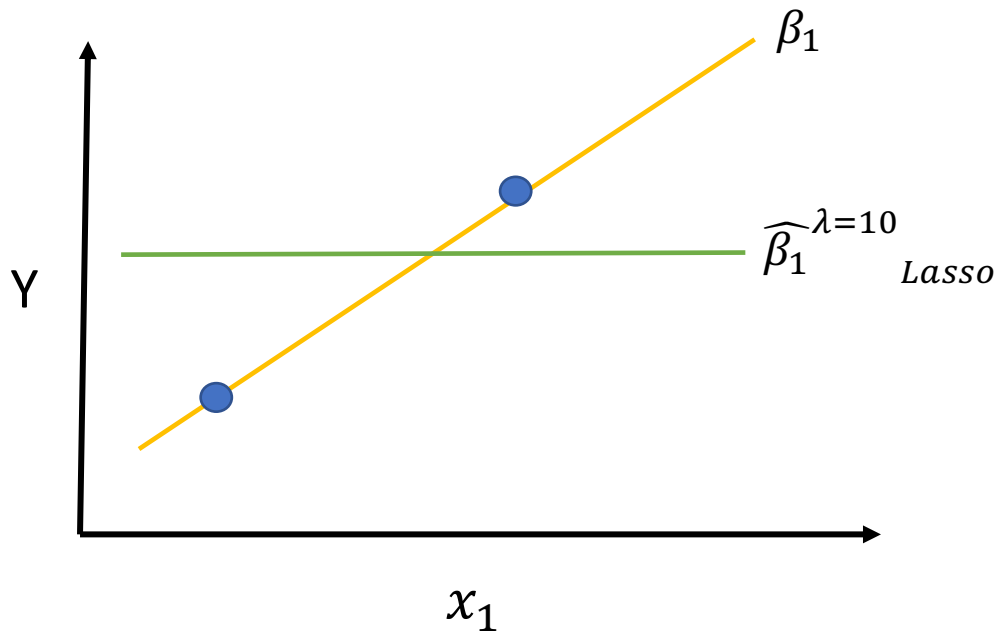
Lasso minimizes the residuals plus lambda times the absolute value of the slope coefficients

Lasso coefficients are still smaller than OLS coefficients

Lasso still accepts a little bias for (hopefully) less variance

Key Lasso Property: Variable Selection

$\hat{\beta}_{Lasso}$ minimizes: *residuals* + $\lambda \cdot (|\beta_1| + |\beta_2|)$



For large values of λ , some slope coefficients will be chosen to be exactly zero

E.g. if we set $\lambda = 10$, maybe $\beta_1^{lasso} = 0$ but $\beta_1^{lasso} \neq 0$

If that happens we effectively remove β_1 from the equation, and we have a variable selection algorithm

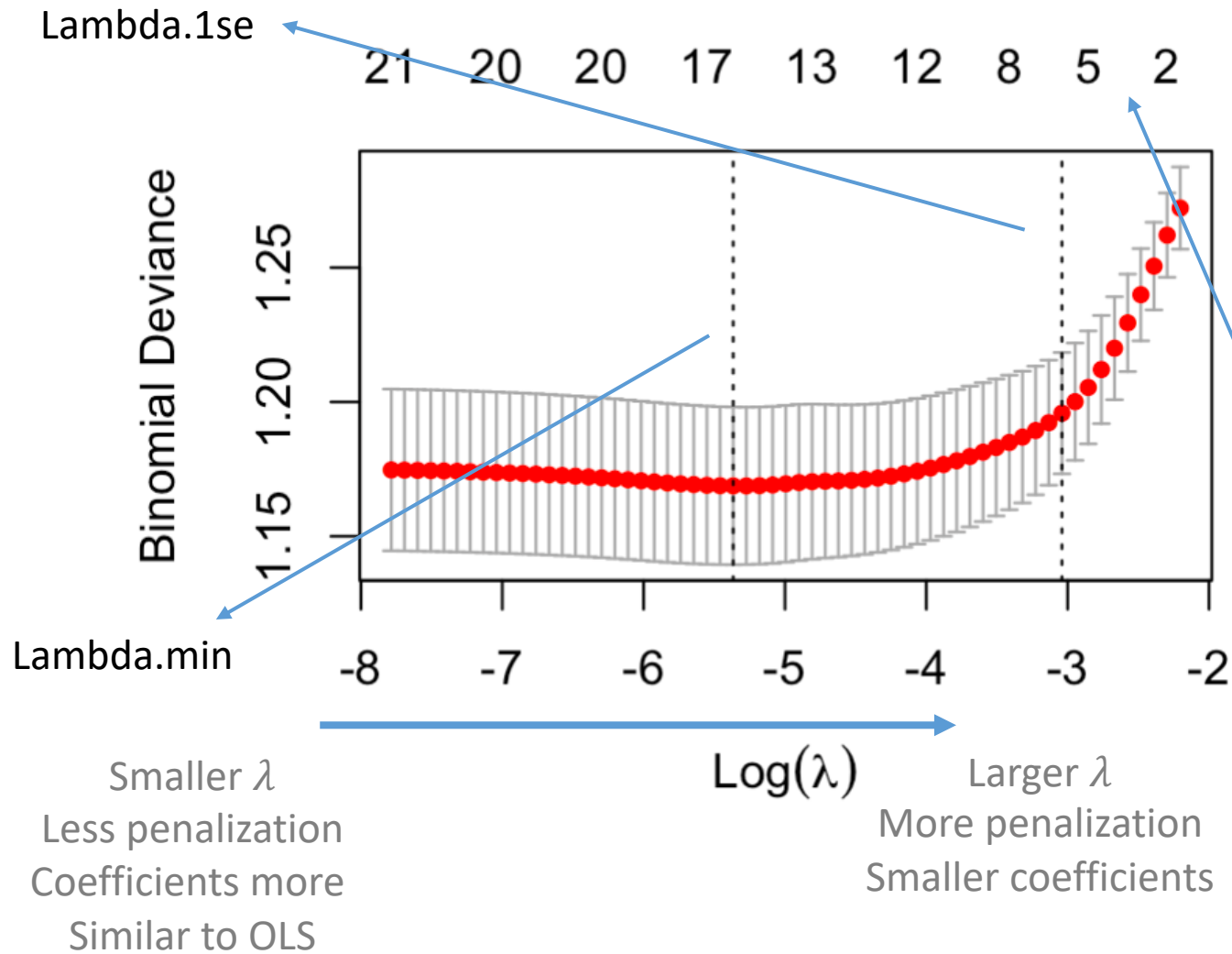
Lasso Model with glmnetUtils

```
lasso_mod <- cv.glmnet(any_bank_account ~ urban + tenureTypeOwn + outerWallsPoor  
  + toiletPoor + elecGrid + bedrooms + aircon + fridges  
  + micros + washers + stereos + DVDplayers + TVs +  
  cellphones + computers + vehicles + cable +  
  internet + numHHmem + numDep + numChildren,  
  data = LFS_train,  
  weights = Weight,  
  family = "binomial",  
  # note alpha = 1 sets lasso!  
  alpha = 1)
```

```
> print(lasso_mod$lambda.min)  
[1] 0.004671366  
> #  
> print(lasso_mod$lambda.1se)  
[1] 0.04781288
```

- We estimate lasso using `cv.glmnet`
- Here we must set `alpha = 1` to estimate Lasso
- Again we get values of `lambda.min` (minimizes cross-validated MSE) and `lambda.1se` (minimum plus 1 SE)

Lasso Cross-Validated MSE Plot



- Lasso MSE plot is very similar
- Top number indicates number of non-zero coefficients for each value of lambda!
- E.g. at this value of lambda, 5 variables are non-zero
- We still have lambda.1se and lambda.min vertical dashed lines but lambda.1se generally shrinks more variables to exactly zero!

Lasso Coefficient Vector

```
> print(lasso_coefs, n = 23)
```

```
# A tibble: 23 × 3
```

	varnames	lasso_min\$s1	lasso_1se\$s1
	<chr>	<dbl>	<dbl>
1	(Intercept)	-0.896	-0.207
2	urban0	-0.009	0
3	urban1	0	0
4	tenureTypeOwn	0	0
5	outerWallsPoor	-0.156	0
6	toiletPoor	-0.038	0
7	elecGrid	0.232	0
8	bedrooms	0	0
9	aircon	0.416	0
10	fridges	0.224	0.195
11	micros	0	0
12	washers	0.355	0.147
13	stereos	-0.04	0
14	DVDplayers	0.154	0
15	TVs	0	0
16	cellphones	0.829	0.395
17	computers	0.568	0.371
18	vehicles	0.519	0.316
19	cable	0.017	0
20	internet	0.045	0
21	numHHmem	-0.108	0
22	numDep	0	0
23	numChildren	0.091	0

- We can build the lasso coefficient vector as we did for Ridge
- Note the higher the lambda (lambda.1se > lambda.min) the more variables that are “shrunk” to zero
- Lasso sets coefficients = 0 if they do not improve the cross-validated MSE
- Ridge will just shrink these coefficients towards zero but will never set them exactly = 0

Another way to write Lasso

Lasso

$$\min_{\beta} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq s$$

Lasso with two variables

$$\min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \quad \text{subject to} \quad |\beta_1| + |\beta_2| \leq s$$

In other words: I give you s as a budget (like setting some lambda)

You can increase your coefficients but the sum of the absolute value of them must be less than s

Another way to write Ridge

Ridge

$$\min_{\beta} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^p (\beta_j)^2 \leq s$$

Ridge with two variables

$$\min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \quad \text{subject to} \quad (\beta_1)^2 + (\beta_2)^2 \leq s$$

In other words: I give you s as a budget (like setting some lambda)

You can increase your coefficients but the sum of the absolute value of them must be less than s

Ridge Versus Lasso Penalty

**Ridge
penalty**

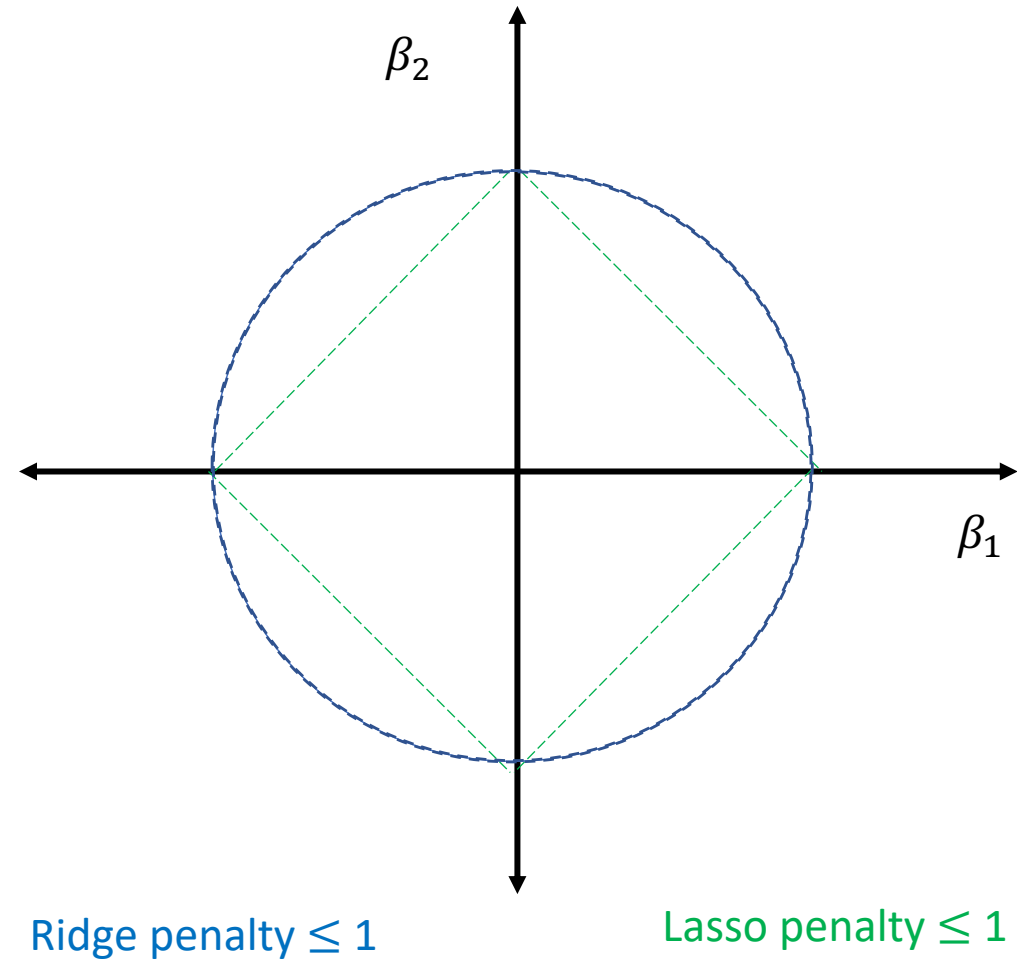
$$(\beta_1)^2 + (\beta_2)^2 \leq 1$$

**Lasso
penalty**

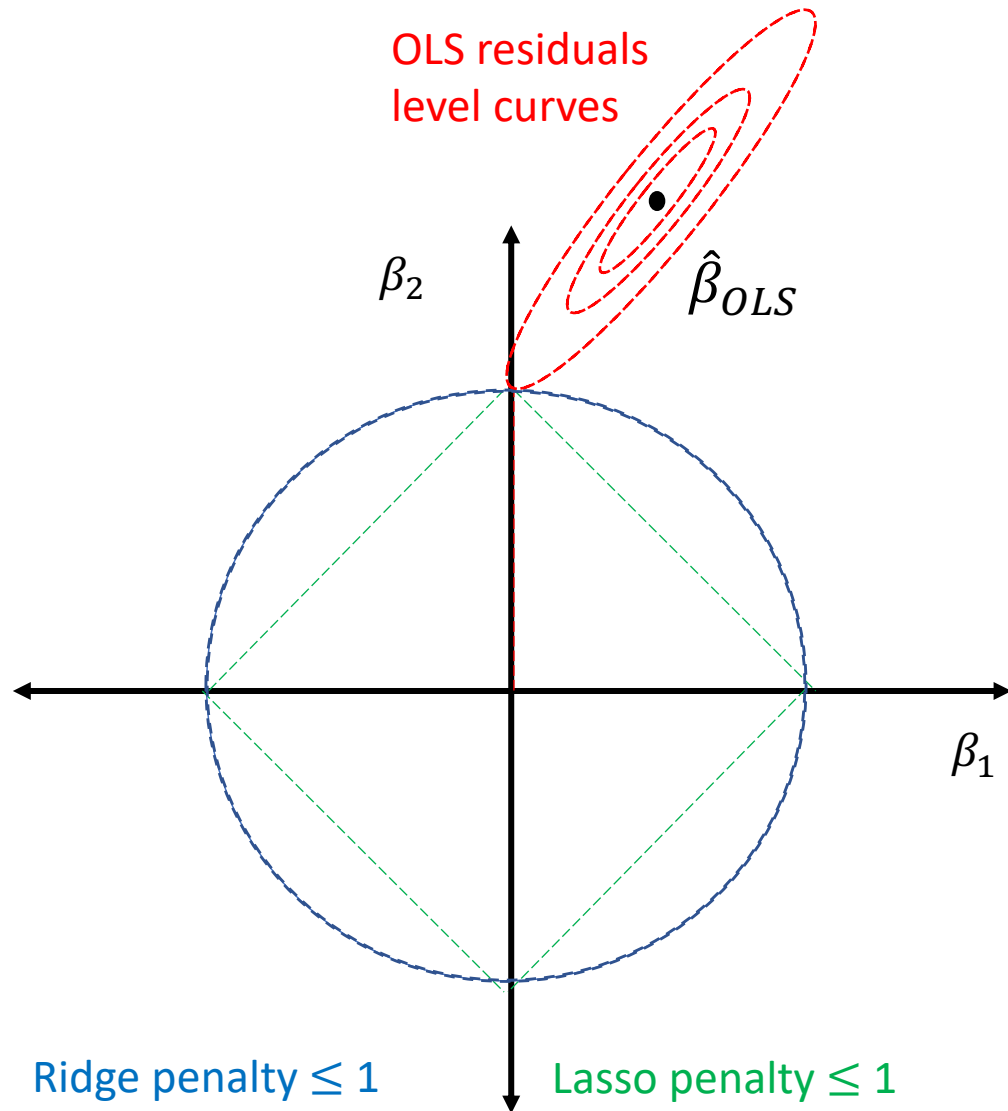
$$|\beta_1| + |\beta_2| \leq 1$$

Let's pick an arbitrary value of $s = 1$

What do these look like graphically?



Ridge and Lasso Equations Redux



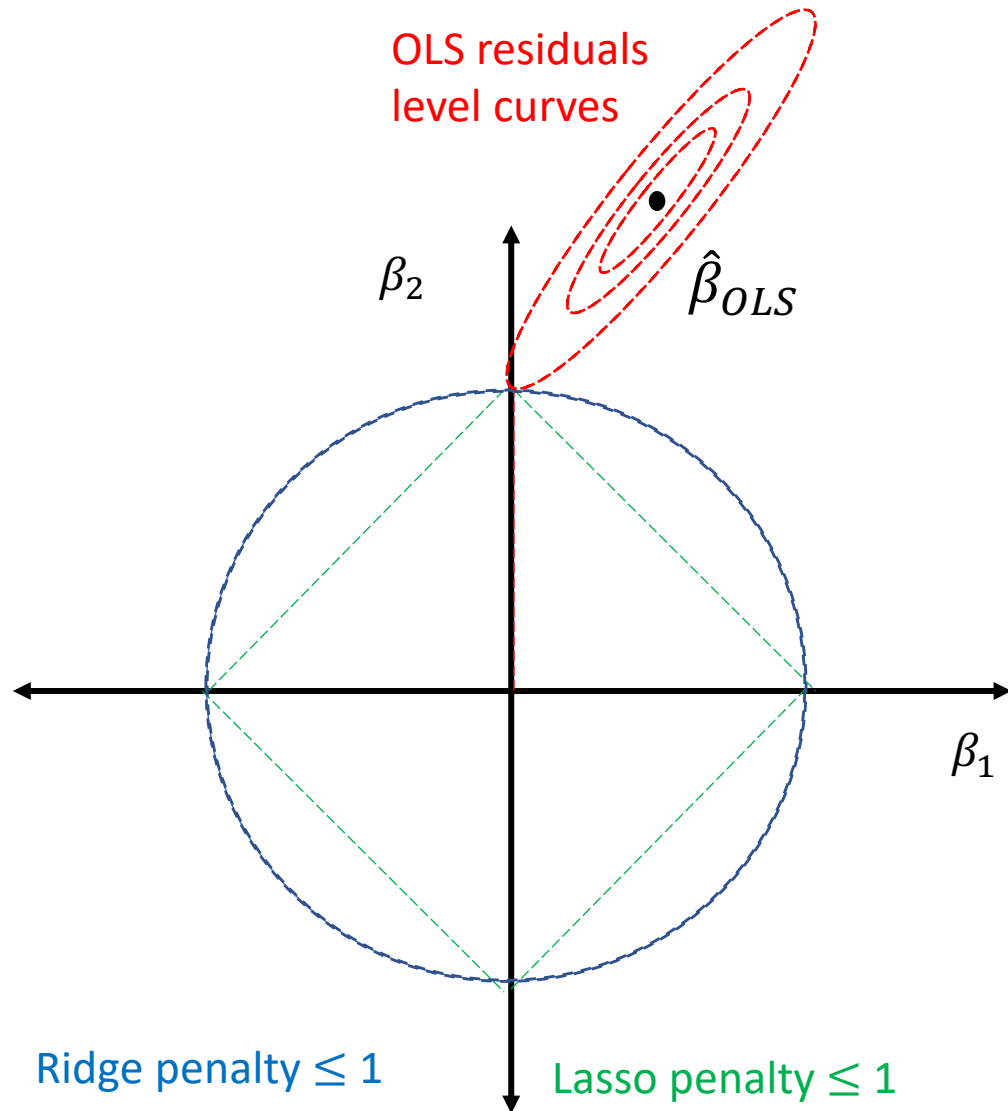
Suppose the optimal OLS beta is this point in black

Meaning, without constraints this point achieves a minimum of the residuals

We can represent that graphically as a series of contour lines where the black dot (OLS beta) is the minimum

Level curves farther from the OLS point are higher residuals

Ridge and Lasso Equations Redux



Graphically what the ridge equation is asking is: “find the lowest residual level curve while staying within the blue circle”

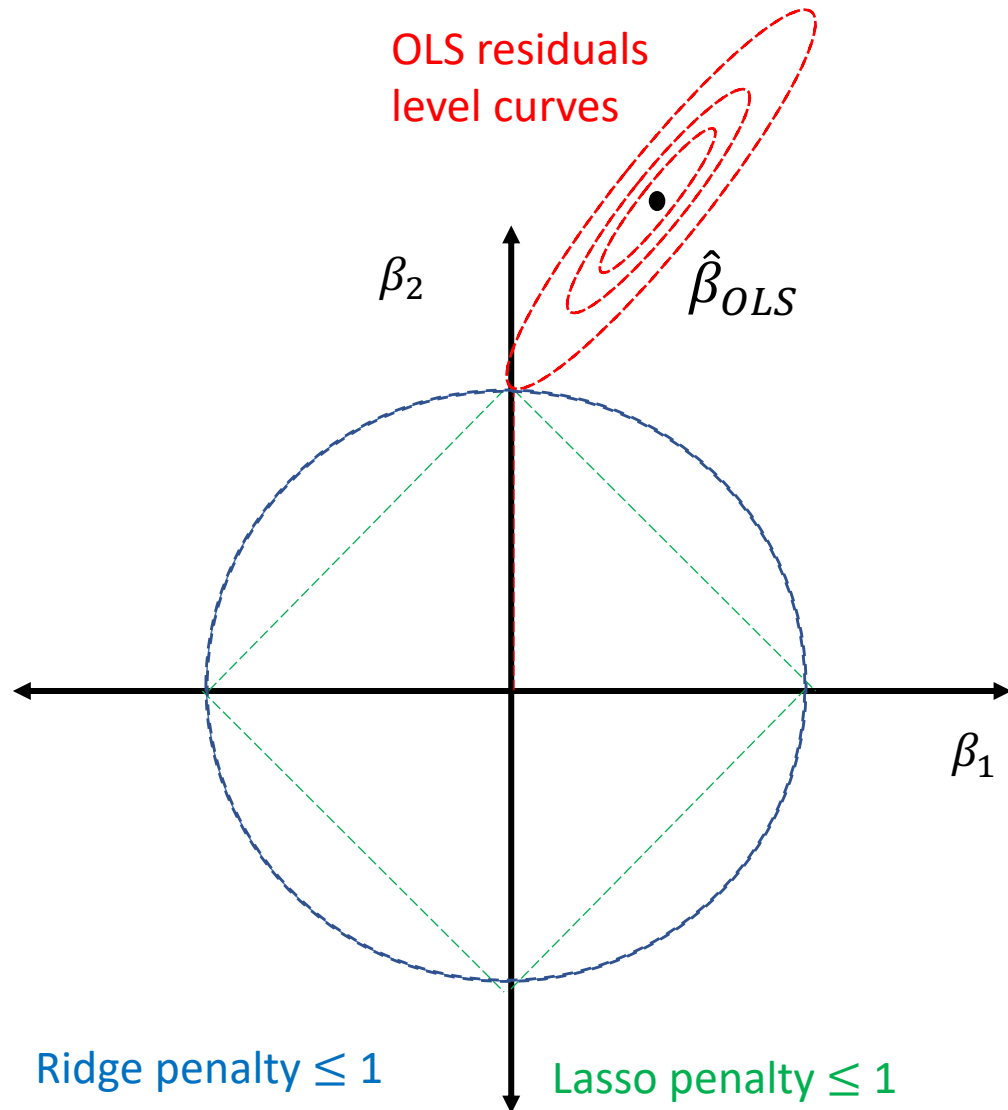
That is the level curve tangent to the blue line

Ridge

$$\min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_{i1} - \beta_1 x_{i2})^2$$

subject to $(\beta_1)^2 + (\beta_2)^2 \leq s$

Ridge and Lasso Equations Redux



Graphically what the Lasso equation is asking is: “find the lowest residual level curve while staying within the green diamond”

That is the level curve tangent to the **green** line

Lasso

$$\min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_{i1} - \beta_1 x_{i2})^2$$

subject to $|\beta_1| + |\beta_2| \leq s$

Ridge and Lasso Equations Redux

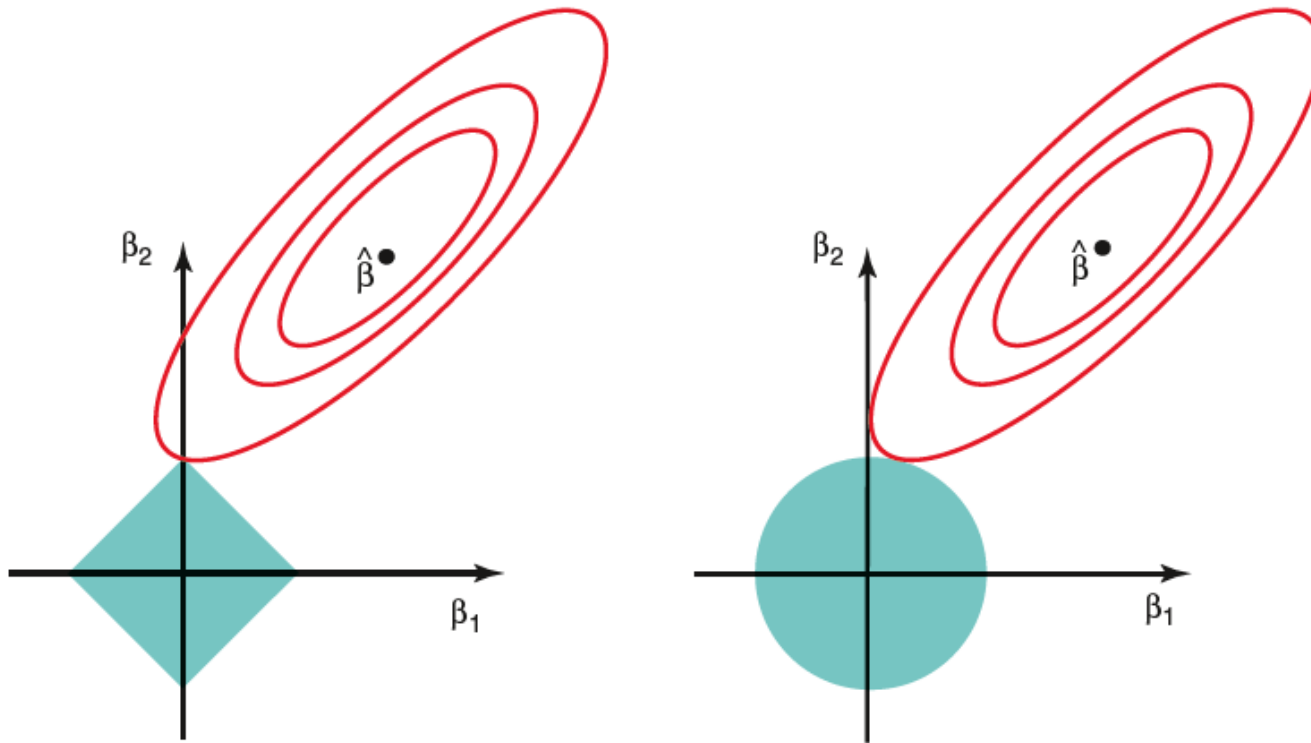


FIGURE 6.7. Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.

Lasso acts as a **variable selector** because the point of tangency for Lasso is often such that one of the variables (here β_1) is zero

Ridge does not have this property, and we see there's still some small value for β_1 in the right plot

Ridge versus Lasso

- Use Lasso when the “data generating process” (DGP, how the data is really formed) is **sparse**
- What is a sparse DGP?
 - Only a few variables really matter!
 - True model is parsimonious
- Ridge should be used when many variables matter a little



ElasticNet



Why Choose? ElasticNet Uses Both Ridge and Lasso Penalty

$$\beta_{ENet} = \min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$
$$+ \lambda \left[\underbrace{\alpha(|\beta_1| + |\beta_2|)}_{\text{Lasso penalty}} + \underbrace{(1 - \alpha)(\beta_1^2 + \beta_2^2)}_{\text{Ridge penalty}} \right]$$

- $\alpha \in [0,1]$ controls the amount of ridge versus lasso penalty
- λ functions as before -> controlling total amount of shrinkage penalty

How to choose λ and α ? Grid Search

$$\beta_{ENet} = \min_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 + \lambda [\alpha (|\beta_1| + |\beta_2|) + (1 - \alpha)(\beta_1^2 + \beta_2^2)]$$

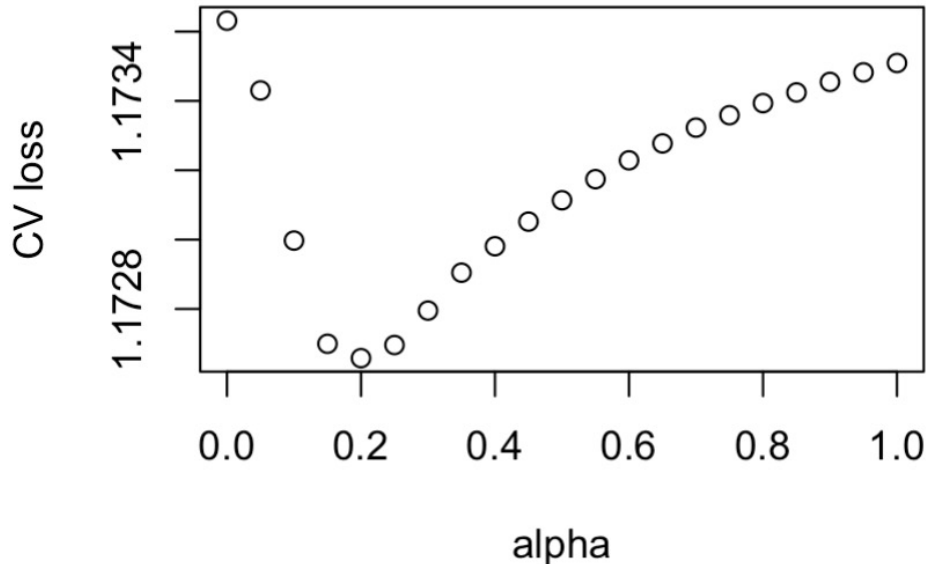
λ	$\alpha = 0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1$
0.5	0.3	0.6	1.3	1.7	4.0
1.0	2.6	3.1	2.1	3.2	4.3
1.5	3.1	3.9	3.2	4.3	5.3
2	3.8	4.6	5.4	6.0	7.4

- We try out a number of different combinations of hyper-parameters
- For each hyper-parameter combination we calculate cross-validated MSE
- Optimal combination has lowest cross-validated MSE

Estimating ElasticNet with `cva.glmnet`

```
enet_mod <- cva.glmnet(any_bank_account ~ urban + tenureTypeOwn + outerWallsPoor  
  + toiletPoor + elecGrid + bedrooms + aircon + fridges  
  + micros + washers + stereos + DVDplayers + TVs +  
  cellphones + computers + vehicles + cable +  
  internet + numHHmem + numDep + numChildren,  
  data = LFS_train, alpha = seq(0,1, by = 0.05),  
  weights = Weight, family = "binomial")
```

```
minlossplot(enet_mod,  
  cv.type = "min")
```



- `cva.glmnet` will estimate a variety of elasticNet models varying alpha from 0 (all ridge) to 1 (all lasso)
- We must specify a sequence of alphas (between zero and 1) to estimate
- The function `minlossplot()` shows us how cross-validated MSE varies as we change alpha
- This plot reveals the minimum alpha value is at $\alpha = 0.2$

Summary – Some Machine Learning Models

- **Parsimony means a model explains equally well using fewer variables**
- **Ridge and Lasso penalize magnitude and number of variables to obtain parsimony and avoid overfitting**
- Ridge uses λ^2 penalization, Lasso uses absolute value penalization
- Cross-validation approximates out of sample fit
- λ key parameter and controls how much parsimony or shrinkage is used
- Higher lambda -> more shrinkage
- Lasso acts as a variable selector
- Use lasso when you believe the true model is sparse/parsimonious/few variables matter
- Use ridge when you believe the true model is not sparse