HUDM 5126
Linear Models and Regression Analysis

Week 10 – Regularization and Dimension Reduction

Shrinkage Methods

- Before we looked at subset selection methods, which can improve fit by dropping variables. The essential idea being that when uninformative variables are used in estimating regression coefficients, the model *overfits* the noise and results in poorer prediction with new cases.
- Another approach to minimizing the impact of noisy predictors is to use *shrinkage methods*, which *shrink* or *regularize* coefficient estimates towards zero.
- In OLS regression, the residual sum of squares is given by

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

- Minimizing *RSS* will yield the normal equations for estimating the betas. Next we will look at two methods that impose an additional restriction on the betas that penalizes large betas.

Ridge Regression

- The essential idea with ridge regression is identical to OLS regression with one exception, there is a penalty added in the form of the l_2 norm of the coefficients.
- The l_2 *norm* is defined as follows. For $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the l_2 norm of $\mathbf{x}$ (also called the Euclidean norm) is

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

- The ridge regression coefficients $\hat{\beta}^R_{j,\lambda}$ are the betas that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\|\beta\|_2$$

- where $\lambda \geq 0$ is a *tuning parameter*.

Ridge Regression

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

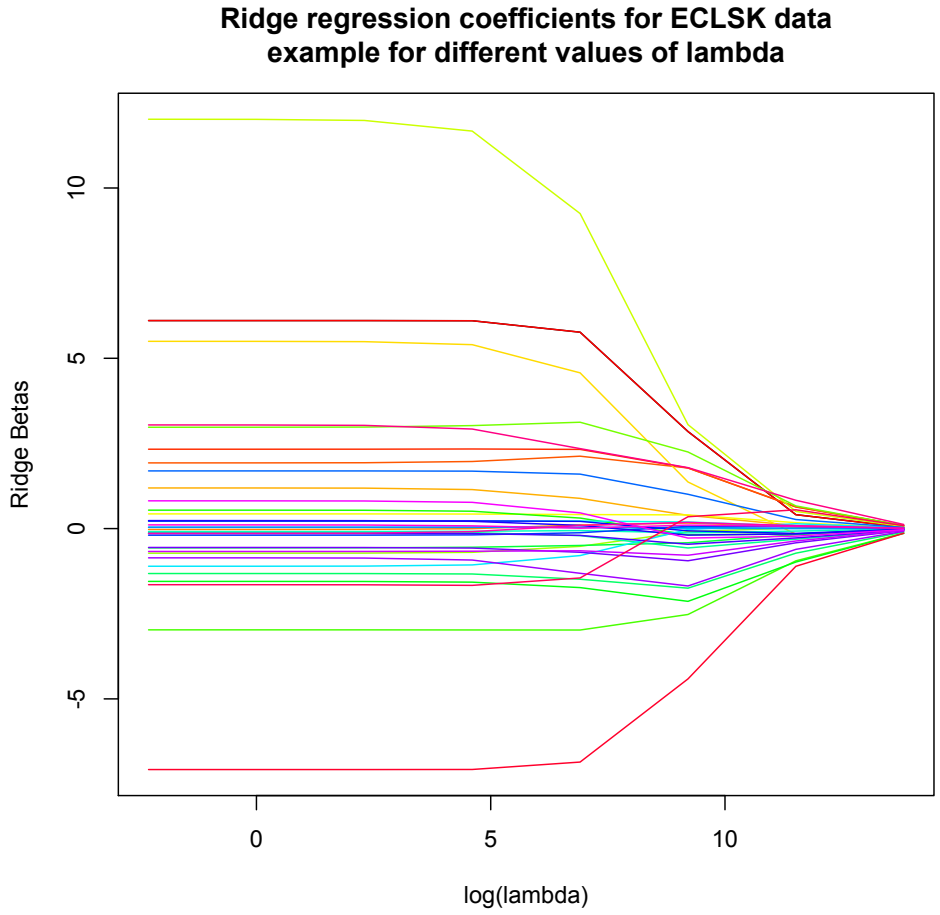
The residual sum of squares is the typical measure of fit used to select the betas. If $\lambda = 0$, the ridge coefficients will be identical to the OLS coefficients. As λ goes to infinity, the coefficients shrink toward zero.

The tuning parameter λ controls the relative contribution of the penalty term to the fit of the regression coefficients. Typically selected by cross-validation.

The shrinkage penalty is small when the betas are near zero. It will be large (and, therefore, penalize more heavily), when one or more betas are large in magnitude.

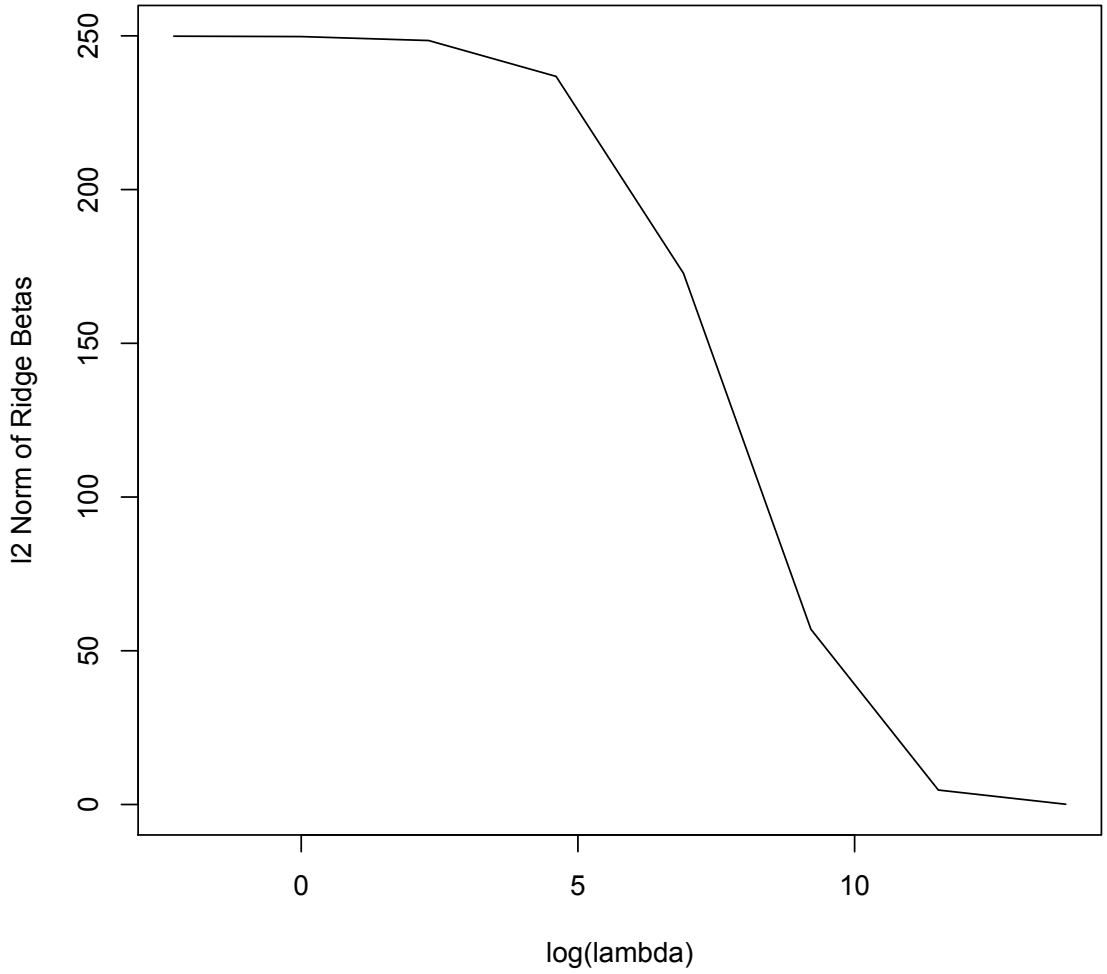
Ridge Regression

- It is also important to note that we do **not** include a penalty for the intercept. The intercept represents the mean of the outcome *Y* when all the predictors are set equal to zero. It does not reflect an association between variables and its magnitude has no bearing on the curviness of the model.



Ridge Regression

- Can also look at the l2 norm of the ridge coefficients for different lambda values.



Ridge Regression Example - ECLSK

	OLS	Ridge		OLS	Ridge
(Intercept)	96.41352370	88.54176805	P1HMAFB	1.16114071	1.20806900
GENDER	3.05365658	2.82143933	WKCAREPK	-0.40414286	-0.26819406
WKWHITE	1.00849025	0.99941291	P1EARLY	0.48985055	0.37188841
WKSESL	1.46858904	1.62544503	wt_ounces	0.73747389	0.75978766
RIRT	-0.30135751	1.38489115	C1FMOTOR	3.29287467	3.05875726
MIRT	12.03212330	8.50612585	C1GMOTOR	-0.34466859	-0.19238713
S2KPUPRI	2.26925739	1.80514609	P1HSCALE	0.17204917	0.15584627
P1EXPECT	0.44589113	0.42540644	P1SADLON	0.08850934	0.02719465
P1FIRKDG	2.16017206	1.57266074	P1IMPULS	-0.09387051	-0.13857066
P1AGEENT	-3.00379619	-2.01862674	P1ATTENI	-0.34379383	-0.44744897
apprchT1	1.90655378	1.99117407	P1SOLVE	-0.49410995	-0.79466611
P1HSEVER	-0.98001336	-0.98024759	P1PRONOU	-0.42500587	-0.41521126
chg14	0.12437526	0.06188985	P1DISABL	0.27210717	0.13312127
P1FSTAMP	-0.51574505	-0.58929915	avg_RIRT	0.64729451	0.07491055
S2KMINOR	-0.81850841	-0.70957604	avg_MIRT	-0.77151458	0.58570494
ONEPARENT	-0.49782480	-0.57543420	avg_SES	1.56521413	1.16524091
STEPPARENT	-0.02681801	-0.05435507	avg_apprchT1	-0.50962275	-0.41918342
P1NUMSIB	-0.09411400	-0.06347336	F5SPECS	-7.07513818	-6.76848026

Ridge Regression

- Coefficient estimates based on OLS regression are *scale equivariant*. That is, if we change one variable by multiplying it by a constant, its OLS regression coefficient will also change (inversely) by that constant.

```
summary(lm(C6R4MSCL ~ MIRT + RIRT, data = eclsk1))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.98428    0.79479  94.345   <2e-16 ***
MIRT         1.52237    0.02867  53.101   <2e-16 ***
RIRT         0.03133    0.02510   1.248    0.212

MIRT2 <- eclsk1$MIRT/100
summary(lm(C6R4MSCL ~ MIRT2 + RIRT, data = eclsk1))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.98428    0.79479  94.345   <2e-16 ***
MIRT2       152.23716    2.86692  53.101   <2e-16 ***
RIRT          0.03133    0.02510   1.248    0.212
```

Another way to say this is that rescaling a predictor by multiplying it by a constant will have no effect on predicted values because

$$X_j \hat{\beta}_j$$

will remain unchanged.

Ridge Regression

- Coefficients estimated by ridge regression, however, are **not** scale equivariant.
- With coefficients estimated by ridge, $X_j\hat{\beta}^R_{j,\lambda}$ can change based on λ, the scaling of the jth predictor, or the scaling of the other predictors.
- Thus, it is important to standardize each predictor by dividing it by its estimated standard deviation before running ridge regression.

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

Ridge Regression

- For OLS, the beta estimates are determined by minimizing the RSS:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2=(y-X\beta)^T(y-X\beta)=\|y-X\beta\|^2$$

- which yields the normal equations: $X^TX\beta-X^Ty=0$
- where,

$$X=\begin{bmatrix}1 & x_{11} & x_{12} & \cdots & x_{1p}\\ 1 & x_{21} & x_{22} & \cdots & x_{2p}\\ 1 & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{np}\end{bmatrix};\beta=\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_p\end{bmatrix};y=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}$$

Ridge Regression

- For ridge regression, the beta estimates are determined by minimizing

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

- where D is a diagonal matrix with a 0 in the leading (1,1) position for the intercept.
- Minimization yields the following: $\left(X^TX+\lambda D\right)\beta-X^Ty=0$

- Note that, because of the addition of λD, it is possible that the system has a unique solution even when the X^TX matrix is not invertible.

Ridge Regression

- Ridge regression can be an improvement over OLS regression because of the relationship between bias and variance of predicted values.
- Assuming the assumptions of OLS are met, the regression coefficient estimates are *unbiased* estimates of the true regression coefficients.
- In ridge regression, this is no longer the case. The coefficients estimated by ridge will carry some bias with them, depending on the value of lambda.
- However, ridge-based estimates will be less variable, and that is where the savings come into play in terms of mean squared prediction error.
- Ridge compares most favorably to OLS in situations where linear regression will likely have a high variance.

Mean Squared Prediction Error

- The mean-squared prediction error is a measure of the average distance between a predictor $\hat{f}$ and the truth f:

$$\text{MSPE} = E\left[\sum_{i=1}^{n}\left(f(x_i) - \hat{f}(x_i)\right)^2\right]$$

- Equivalently, MSPE may be expressed as the sum of squared biases and the sum of variances of predicted values. This decomposition is analogous to the decomposition of MSE into bias squared plus variance.

$$\text{MSPE} = \sum_{i=1}^{n}\left(E\left[\hat{f}(x_i)\right] - f(x_i)\right)^2 + \sum_{i=1}^{n}\text{var}\left[\hat{f}(x_i)\right]$$

Lasso Regression

- Lasso regression is identical to ridge regression with one exception: the *l*1 norm is used instead of the *l*2 norm.
- The *l*1 *norm* is defined as follows. For $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the *l*1 norm of **x** is

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} x_i$$

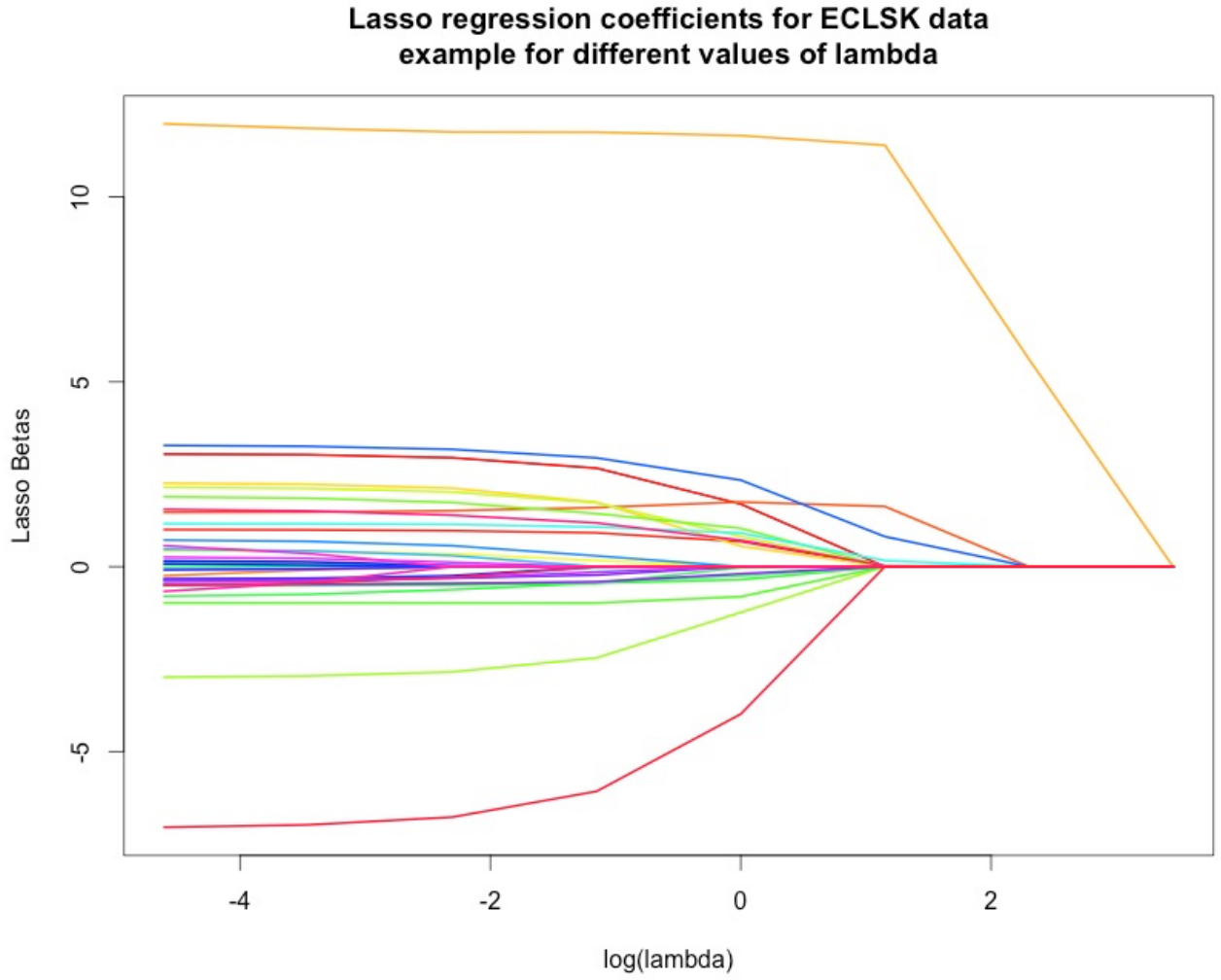
- The ridge regression coefficients are the betas that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\|\beta\|_1$$

- where $\lambda \geq 0$ is the *tuning parameter*.

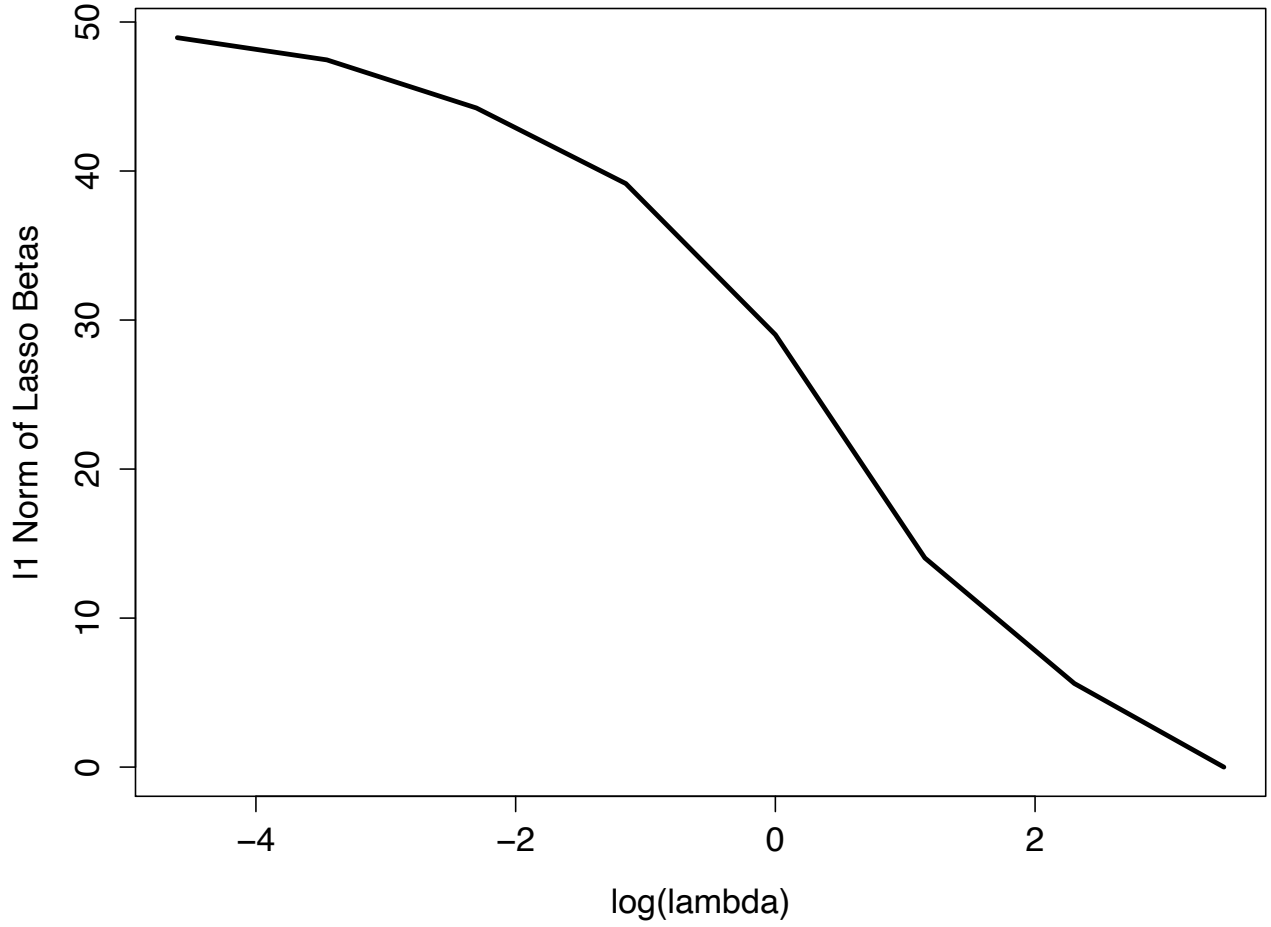
Lasso Regression Example - ECLSK

- The essential difference between lasso and ridge regression is that lasso can set some coefficients to be exactly zero.



Ridge Regression Example - ECLSK

- Can also look at the *l*2 norm of the ridge coefficients for different lambda values.



Lasso Regression Example - ECLSK

	OLS	Ridge	Lasso
(Intercept)	96.41	86.30	89.11
GENDER	3.05	2.82	2.40
WKWHITE	1.01	1.00	0.84
WKSESL	1.47	1.63	1.66
RIRT	-0.30	1.38	0.00
MIRT	12.03	8.51	11.74
S2KPUPRI	2.27	1.81	1.42
P1EXPECT	0.45	0.43	0.01
P1FIRKDG	2.16	1.57	1.48
P1AGEENT	-3.00	-2.02	-2.14
apprchT1	1.91	1.99	1.33
P1HSEVER	-0.98	-0.98	-0.94
chg14	0.12	0.06	0.00
P1FSTAMP	-0.52	-0.59	-0.42
S2KMINOR	-0.82	-0.71	-0.32
ONEPARENT	-0.50	-0.58	-0.38
STEPPARENT	-0.03	-0.05	0.00
P1NUMSIB	-0.09	-0.06	0.00

	OLS	Ridge	Lasso
P1HMAFB	1.16	1.21	1.03
WKCAREPK	-0.40	-0.27	0.00
P1EARLY	0.49	0.37	0.00
wt_ounces	0.74	0.76	0.19
C1FMOTOR	3.29	3.06	2.79
C1GMOTOR	-0.34	-0.19	0.00
P1HSCALE	0.17	0.16	0.00
P1SADLON	0.09	0.03	0.00
P1IMPULS	-0.09	-0.14	0.00
P1ATTENI	-0.34	-0.45	-0.16
P1SOLVE	-0.49	-0.79	-0.38
P1PRONOU	-0.43	-0.42	-0.05
P1DISABL	0.27	0.13	0.00
avg_RIRT	0.65	0.07	0.00
avg_MIRT	-0.77	0.59	0.00
avg_SES	1.57	1.17	1.07
avg_apprchT1	-0.51	-0.42	0.00
F5SPECS	-7.08	-6.77	-5.55

Ridge vs. Lasso

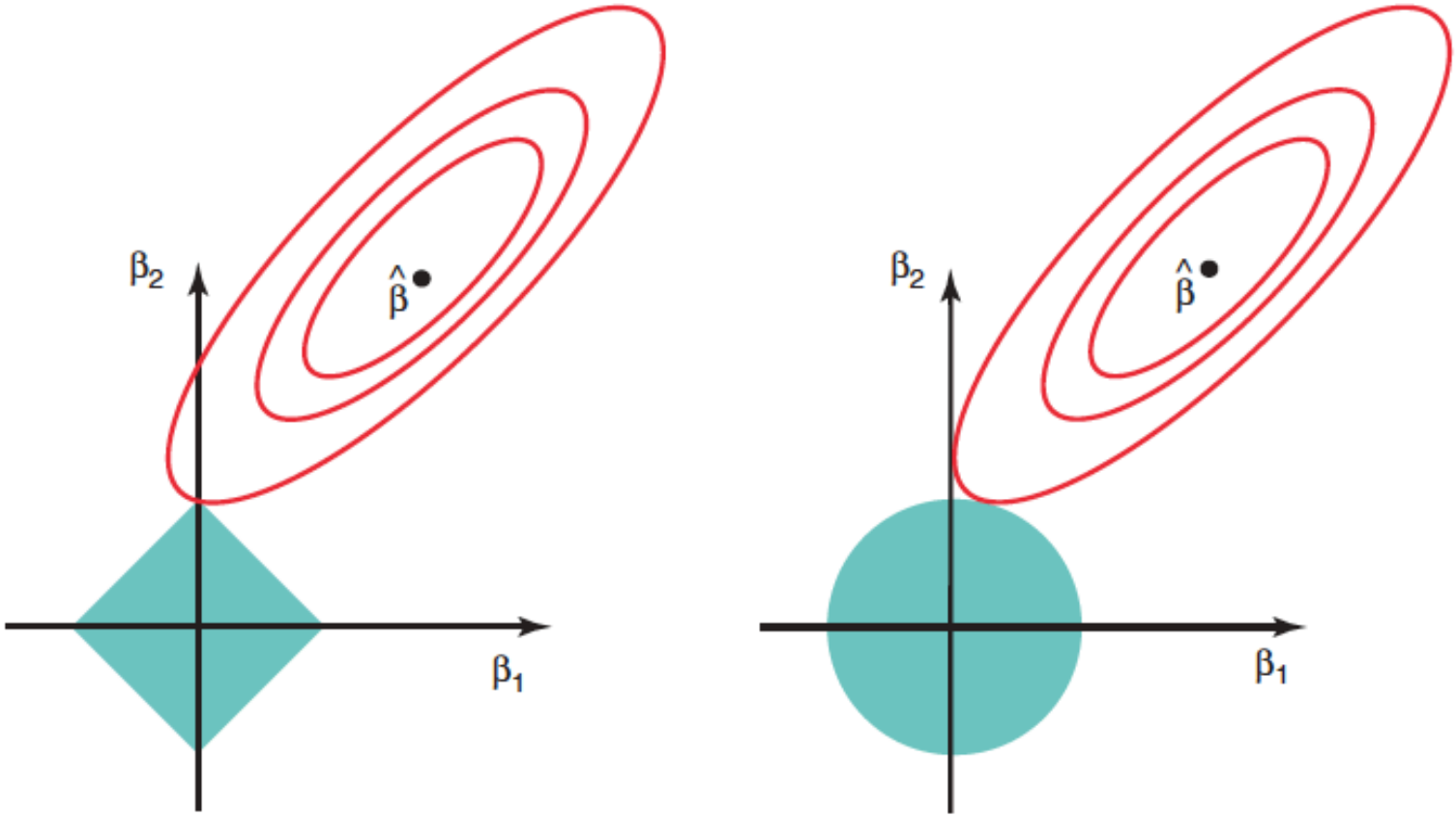


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Dimension Reduction

- All of the model selection and regularization methods we have considered so far control the variance of the prediction errors in one of two ways.
  1. The subset selection methods select a subset of the predictors.
  2. The regularization methods attenuate regression coefficients by shrinking them to be closer to zero.
- Thus, subset selection and regularization methods work with the original predictors themselves, $X_1, X_2, \ldots, X_p$.
- In contrast, *dimension reduction* methods transform the predictors and then fit a model on the transformed variables.

Dimension Reduction

- Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ linear combinations of the original p predictors. In other words, for some constants $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$, $m = 1, \ldots, M$,

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

- We then can fit a typical linear regression via OLS of the outcome on the *transformed* predictors as

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

- This is an example of dimension reduction because now we use $M + 1$ total coefficients $\theta_0, \theta_1, \ldots, \theta_M$ instead of $p + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_p$.

Dimension Reduction

- By combining the equation on the previous slide, one can see that

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

- where

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

- Thus, dimension reduction places a constraint on the regression coefficients, which renders them biased.
- However, especially in cases where p is large in relation to n, dimension reduction can reduce the variance of the coefficients by an amount that offsets the bias and, therefore, has lower MSPE.

Principal Components Analysis

Principal components analysis is a method which computes new axes for multivariate data so that

1. the first axis explains the **most possible variation** in the data,
2. the second axis explains the most possible **remaining** variation **and** is orthogonal to the first axis,
3. the third axis explains the most possible **remaining** variation **and** is orthogonal to the first and second axes,
4. and so on until all the variables have been accounted for.

Principal Components Regression

- The goal of principal components analysis is often the description of data using a smaller number of variables than in the original data matrix.
- In PCA, weighted linear combinations of predictors are used to define new *principal component* vectors.

$$PC_{(1)} = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p$$

$$PC_{(2)} = a_{12}X_1 + a_{22}X_2 + \cdots + a_{p2}X_p$$

$$\cdots$$

$$PC_{(p)} = a_{1p}X_1 + a_{2p}X_2 + \cdots + a_{pp}X_p$$

Principal Components Analysis (PCA)

- These weighted linear combinations are unique in that,
  - subject to the constraint that the sum of the squared weights a_{ij} equal 1, they maximize the variance among the objects being measured,
  - the components are ordered so that the first accounts for the largest amount of variance among the objects, the second for the next largest amount of variance, and so on.
  - all of the components are orthogonal, or uncorrelated so that $\sum_{j=1}^{p} a_{ij} a_{ik} = 0$.
- In other words, there is no redundancy or overlap in the variance explained by each principal component.