Ch6 - Regularization

Contents

Textbook: James et al. ISLR 2ed.

6A Subsection

[ToC]

A.1 Subset Selection

- Non-linear, but still additive relationship
- Linear model is surprisingly competitive in real world modeling.
- Improve linear model by not using Least Square method.
- Variable Selection (P-values add up)

  1. Subset Selection
  2. Shrinkage
  3. Dimension Reduction

A.2 Best Subset Selection Algorithm

1. Let M_0 denote the null model (no predictors).
2. For $k = 1, 2, \ldots, p$
   (a) Fit all $\binom{p}{k}$ models with k predictors.
   (b) Pick the best (min RSS, or max R^2), call it M_k
3. Pick the best among $M_0, \ldots, M_p$, by using C-V prediction error, AIC, BIC, or adjusted R^2.

A.3 RSS or SSE?

	Naming Convention	1	2	3	4	5
A	$\sum(Y_i - \bar{Y})^2$	TSS	$TotSS$	$TotSS$	$SSTot$	$SSTot$
B	$\sum(Y_i - \hat{Y}_i)^2$	SSE	$ErrSS$	$ResSS$	$SSRes$	$SSErr$
C	$\sum(\hat{Y}_i - \bar{Y})^2$	SSR	$RegSS$	$ExpSS$	$SSReg$	$SSReg$

```
A. Total Sum of Squares (TSS)

B. Sum of Squared Estimate of Errors (SSE)
   Error Sum of Squares (ESS)
   Residual Sum of Squares (RSS) (James ISLR)
   Sum of Squared Residuals (SSR)


C. Sum of Squares (due to) Regression (SSR)
   Regression Sum of Squares (RSS)
   Explained Sum of Squares (ESS)
   Model Sum of Squares (MSS)
```

A.4 Best Subset

1. p=10, you need to try 1000 models. p=20, try 1000000 models.

2. not feasible if $p > 40$

A.5 Need Global Measure of Fit

1. RSS always decrease with extra variable, and R^2 always decrease.

2. Need some measure of fit that can be used to compare p=1 vs p=10.

3. AIC and BIC (under regression model with Gaussian error)

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2p\hat{\sigma}^2), \qquad \text{BIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + \log(n)p\hat{\sigma}^2)$$

   where $\hat{\sigma}^2$ is full model MSE.

4. Adjusted-R^2, Av validation MSE from CV

A.6 Forward Selection

1. M_0 is the model with no predictor.

2. For $k = 0, \ldots, p-1$

   (a) Consider all $p-k$ models that augment the predictors in M_k with one additional predictor.

   (b) Choose the best among these $p-k$ models, and call it M_{k+1}. (min RSS or max R^2).

3. Select a best among M_0, ..., M_p using Av. validation CV MSE, AIC, BIC, or adjusted R^2.

A.7 Backward Selection

1. M_p is the full model with all predictors.
2. For $k = p, p - 1 \ldots, 1$
   (a) Consider all k models that contain all but one of the predictors in M_k.
   (b) Choose the best among these k models, and call it M_{k-1}. (min RSS or max R^2).
3. Select the best among $M_0, ..., M_p$ using Av. validation CV MSE, AIC, BIC, or adjusted R^2.

A.8 Best Model?

Model	# of Non Intercept Parameters	Parameters	R^2	AIC
1	0	I	0	1.9
2	1	I, 1	0.56	1.4
3	1	I, 2	0.57	1.2
4	1	I, 3	0.55	1.6
5	1	I, 4	0.52	1.7
6	1	I, 5	0.51	1.8
7	2	I, 1, 2	0.61	1.0
8	2	I, 1, 3	0.64	0.5
9	2	I, 1, 4	0.63	0.8
10	2	I, 1, 5	0.69	0.0
11	2	I, 2, 3	0.61	1.0
12	2	I, 2, 4	0.62	0.9
13	2	I, 2, 5	0.68	0.2
14	2	I, 3, 4	0.66	0.4
15	2	I, 3, 5	0.64	0.5
16	2	I, 4, 5	0.60	1.1

Model	# of Non Intercept Parameters	Parameters	R^2	AIC
17	3	I, 1, 2, 3	0.73	1.3
18	3	I, 1, 2, 4	0.71	1.5
19	3	I, 1, 2, 5	0.72	1.4
20	3	I, 1, 3, 4	0.75	1.0
21	3	I, 1, 3, 5	0.76	0.8
22	3	I, 1, 4, 5	0.79	0.2
23	3	I, 2, 3, 4	0.78	0.6
24	3	I, 2, 3, 5	0.74	1.2
25	3	I, 2, 4, 5	0.75	1.1
26	3	I, 3, 4, 5	0.73	1.3
27	4	I, 1, 2, 3, 4	0.88	1.6
28	4	I, 1, 2, 3, 5	0.80	2.1
29	4	I, 1, 2, 4, 5	0.87	1.8
30	4	I, 1, 3, 4, 5	0.83	2.0
31	4	I, 2, 3, 4, 5	0.85	1.9
32	5	I, 1, 2, 3, 4, 5	0.90	3.5

Try Best Subset, Forward, and Backward selection.

A.9 Shrinkage

- Ordinary Least Squares

$$\text{RSS} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \quad \text{where} \quad \hat{Y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.$$

- Ridge Regression

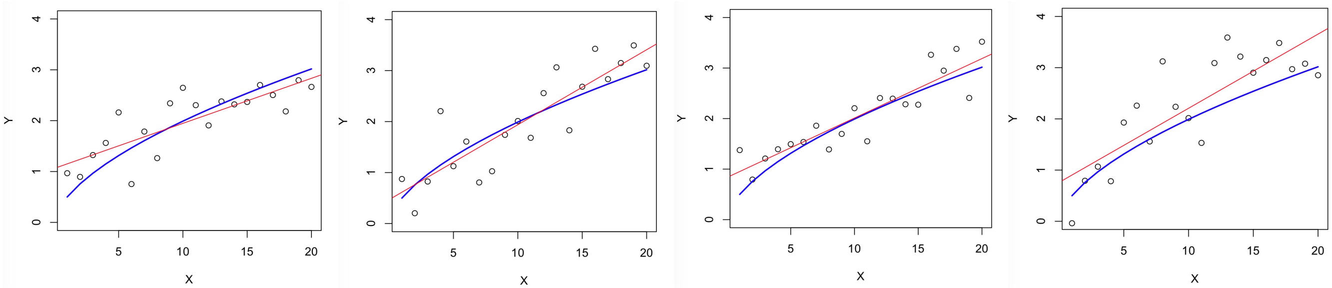
$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

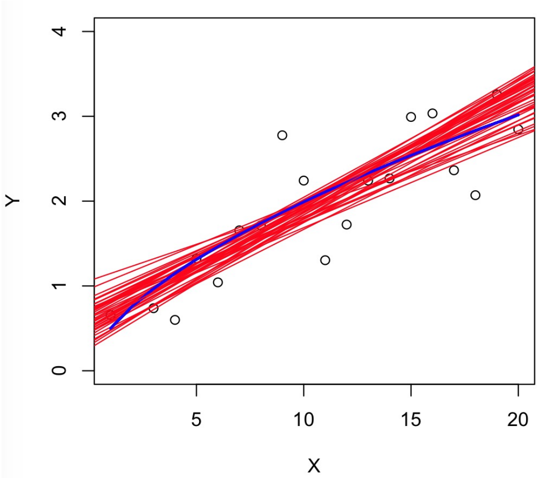
- Lasso Regression

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

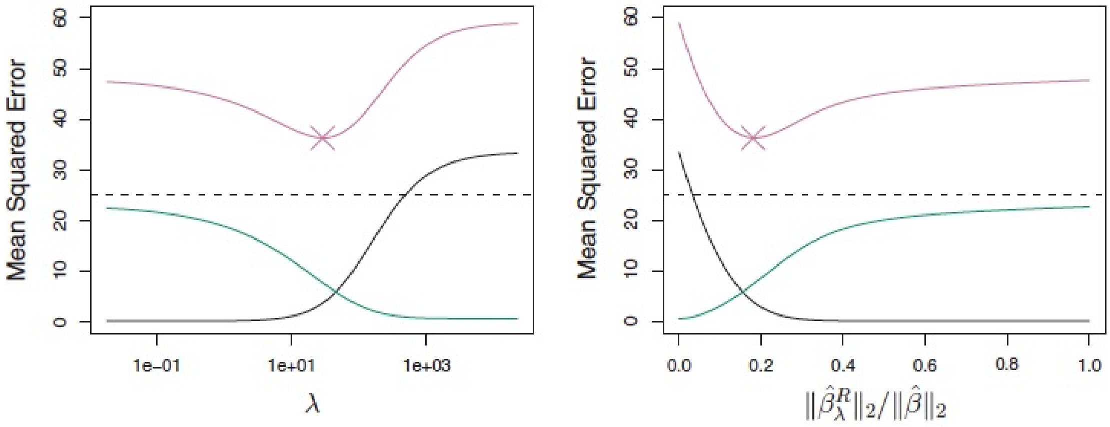
- Tuning parameter λ
- Shrinkage penalty (does not include β_0)
- Use CV to choose best Tuning parameter (Av validation MSE)
- OLS estimators are scale invariant
- Shrinkage penalty is not. So predictors **must be standardized**.

A.10 Bias-Variance Trade off



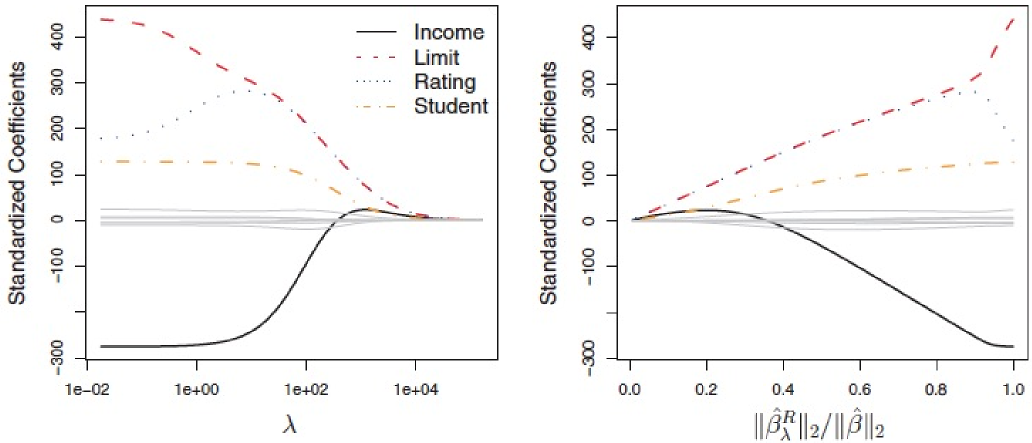


0
1
2
3
4
Y
5
10
15
20
X



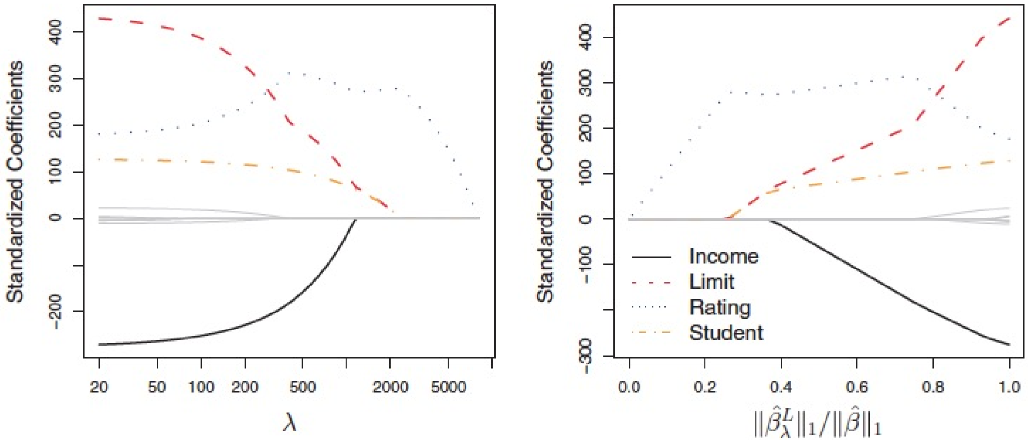
Black: Sq bias, Green: Variance.

A.11 Ridge



Income
Limit
Rating
Student
Standardized Coefficients
400
300
200
100
0
−100
−300
1e−02
1e+00
1e+02
1e+04
λ
Standardized Coefficients
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$

A.12 Lasso


Standardized Coefficients
400
300
200
100
0
−200
20
50
100
200
500
2000
5000
λ
Standardized Coefficients
400
300
200
100
0
−100
−300
Income
Limit
Rating
Student
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$

A.13 Another formulation

- Ordinary Least Squares

$$\text{RSS} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

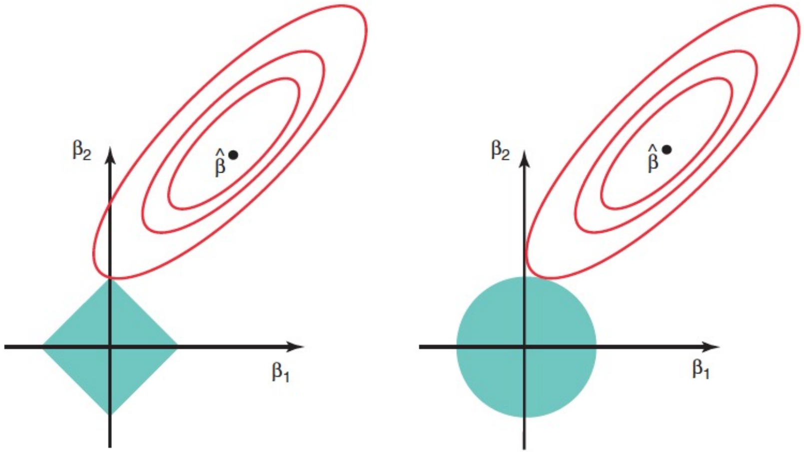
- Ridge Regression

$$\min_\beta RSS \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

- Lasso Regression

$$\min_\beta RSS \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

- βs on a budjet



β_2
$\hat{\beta}$
β_1
β_2
$\hat{\beta}$
β_1

A.14 Boston Data

OLS

```
Call:
lm(formula = medv ~ ., data = Train.set)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.007098   0.026689   0.266  0.79042
crim        -0.064499   0.040993  -1.573  0.11644
zn           0.123939   0.039960   3.102  0.00207 **
indus        0.024692   0.052226   0.473  0.63663
chas         0.071541   0.027416   2.609  0.00942 **
nox         -0.261382   0.059066  -4.425 1.25e-05 ***
rm           0.263032   0.035818   7.344 1.25e-12 ***
age          0.033274   0.047801   0.696  0.48679
dis         -0.338757   0.052623  -6.437 3.61e-10 ***
rad          0.304885   0.072104   4.228 2.94e-05 ***
tax         -0.238095   0.077901  -3.056  0.00240 **
ptratio     -0.227566   0.035880  -6.342 6.33e-10 ***
black        0.078829   0.031817   2.478  0.01365 *
lstat       -0.454766   0.047574  -9.559  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.532 on 386 degrees of freedom
Multiple R-squared:  0.7326,Adjusted R-squared:  0.7236
F-statistic: 81.33 on 13 and 386 DF,  p-value: < 2.2e-16
```

LASSO

```
CV.for.lambda$lambda.min
[1] 0.002343072

> FitLasso <- glmnet(x, y, alpha = 1, lambda = CV.for.lambda$lambda.min)
>coef(FitLasso)
14 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  0.007387128
crim        -0.057107142
zn           0.113845718
indus        0.002936396
chas         0.071900326
nox         -0.241072278
rm           0.266677208
age          0.022894864
dis         -0.330627174
rad          0.262872998
tax         -0.198488772
ptratio     -0.223658142
black        0.076838064
lstat       -0.449521122
```

Ridge

```
> CV.for.lambda$lambda.min
[1] 0.07496112
> FitRidge <- glmnet(x, y, alpha = 0, lambda = CV.for.lambda$lambda.min)
> coef(FitRidge)
14 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  0.008902216
crim        -0.050744039
zn           0.088435303
indus       -0.024351436
chas         0.075542575
nox         -0.171954704
rm           0.282646332
age          0.009136308
dis         -0.256595401
rad          0.155320997
tax         -0.117043223
ptratio     -0.211067354
black        0.079910836
lstat       -0.398769344
```

A.15 Test MSE

```
> OLS
          RMSE   Rsquare
medv 0.4414927 0.7793783
```

```
> LASSO
          RMSE  Rsquare
medv 0.4408182 0.780233
```

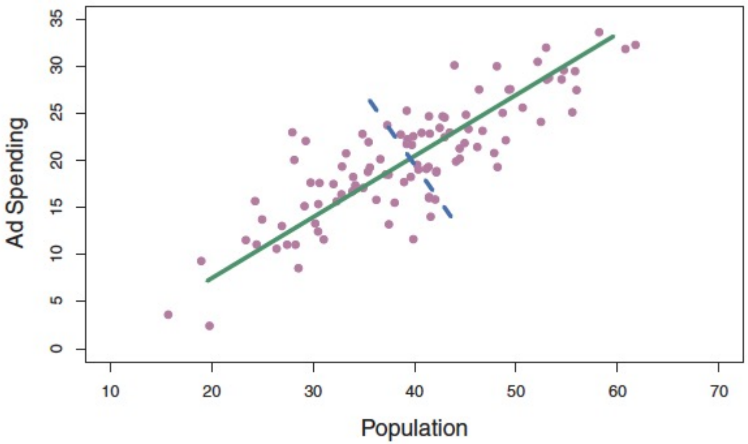
```
> RIDGE
          RMSE   Rsquare
medv 0.4368946 0.7873559
```

A.16 Ridge vs Lasso

- Lasso can be used as dimention reduction tool.
- Lasso model is easier to interpret
- If no coefficients were suppressed by Lasso, then Ridge is better.

A.17 Dimention Reduction

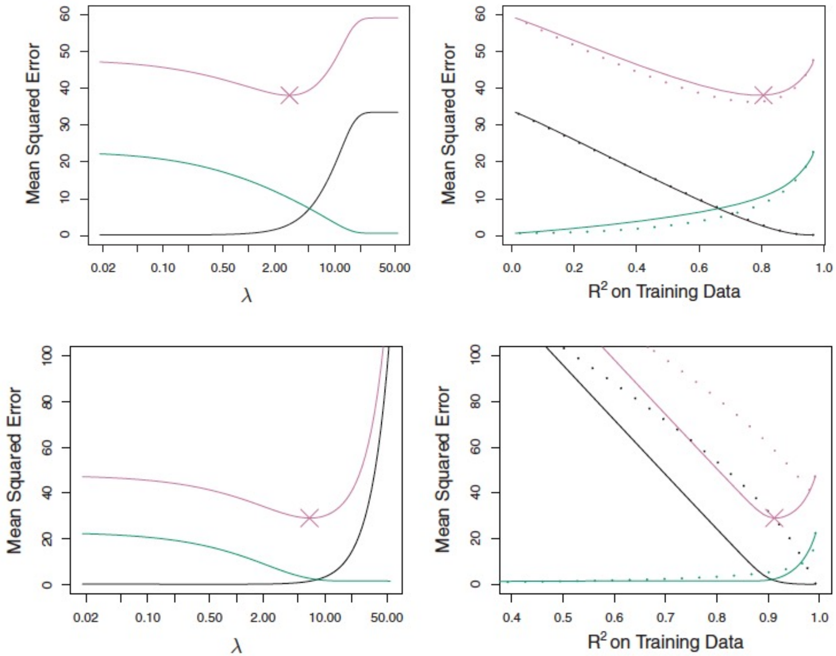
Principal Component Regression



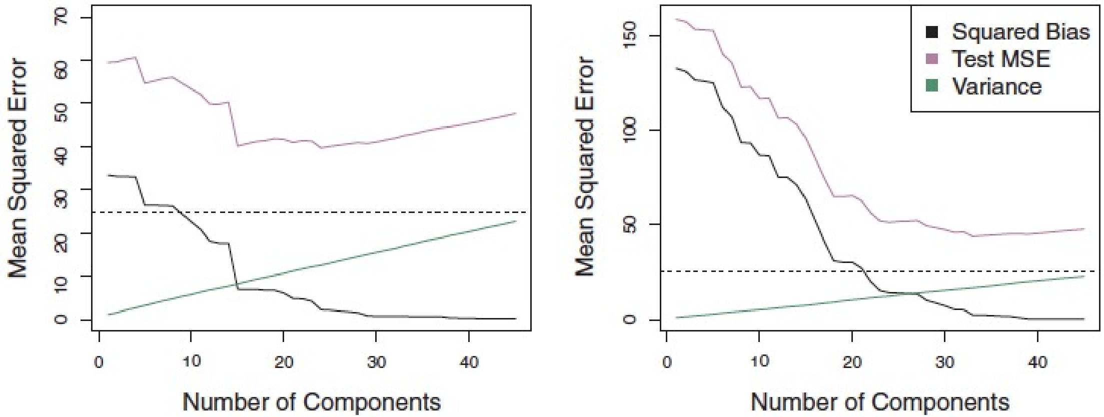
Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

A.18 Principal Component Regression

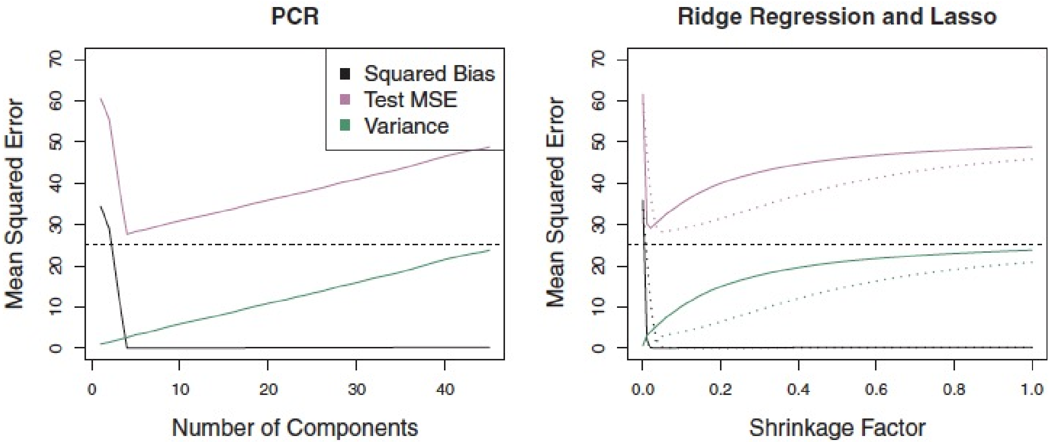
Lasso, and Lasso + Ridge



Mean Squared Error
0 10 20 30 40 50 60
0.02 0.10 0.50 2.00 10.00 50.00
λ
Mean Squared Error
0 10 20 30 40 50 60
0.0 0.2 0.4 0.6 0.8 1.0
R2 on Training Data
Mean Squared Error
0 20 40 60 80 100
0.02 0.10 0.50 2.00 10.00 50.00
λ
Mean Squared Error
0 20 40 60 80 100
0.4 0.5 0.6 0.7 0.8 0.9 1.0
R2 on Training Data



When only 5 predictor is related to response


PCR
Squared Bias
Test MSE
Variance
Mean Squared Error
0
10
20
30
40
50
60
70
0
10
20
30
40
Number of Components
Ridge Regression and Lasso
Mean Squared Error
0
10
20
30
40
50
60
70
0.0
0.2
0.4
0.6
0.8
1.0
Shrinkage Factor