STATS762 Regression for Data Science

Regularized regression

May 15, 2019

Variable selection problem - Revision

In linear regression, the more independent variables is better?
NO! If there are too many independent variables, it will be overfitted.

- Overfitted models describe random error or noise instead of any underlying relationship.
- Poor predictive performance on test data

If there are not enough independent variables, it will be under-fitted.

- There is not enough information in independent variables to find relationship to the dependent variable.
- Poor predictive performance on test data.

Hence carefully selected features improves model predictability.

Variable selection problem - Revision

Given a p-dimensional covariate, $X \in \mathbb{R}^p$, there are 2^p possible combinations of X; 2^p possible models.
Some well known variable selection methods are;

(a) Criterion - Model is chosen based on criterion. For example AIC, BIC, et al. The best subset depends on the complexity penalty of criterion.

(b) Forward-Stepwise Selection - Start with a null model and add significant covariates sequentially until a model contains all significant covariates.

(c) Backward-Stepwise Selection

High-dimensional data

Difficulties associated with a large p;

- Collinearity of covariates - Covariates are linearly dependent and it is impossible to separate the contribution of the individual covariates.
- Singularity of $X^T X$ - When $rank(X) < p$, there are infinitely many solutions in the least squares problem

$$i.e., \hat{\beta}^{OLS} = (X^T X)^{-1} X^T Y .$$

  This type of nonuniqueness makes interpretation of solutions meaningless.

- Even if $rank(X) = p$ (the unique least squares solution exists), the solution will be poor if p is moderately close to n.

When $p > n$ (called a wide data), most of existing estimator fails.

High-dimensional data

When p is high, the OLS often fails and those subset selection methods (for the variable selection problem) are not useful.

How to find the best subset when many potential aggressors (large p)?

We need stronger penalty on model complexity (model size).

Ozone data

The data contains 179 measurements of ozone concentration in the atmosphere. 7 main effects which, jointly with the quadratic terms and second order interactions, produce the above-mentioned $p = 35$ possible regressors.[1]

- y Response = Daily maximum 1-hour-average ozone reading (ppm) at Upland, CA
- x4 500-millibar pressure height (m) measured at Vandenberg AFB
- x5 Wind speed (mph) at Los Angeles International Airport (LAX)
- x6 Humidity (percentage) at LAX
- x7 Temperature (Fahrenheit degrees) measured at Sandburg, CA
- x8 Inversion base height (feet) at LAX
- x9 Pressure gradient (mm Hg) from LAX to Daggett, CA
- x10 Visibility (miles) measured at LAX

The rest 27 covariates are quadratic terms. For example, x4.x5=x4*x5

[1] Berger, J. and Molina, G. (2005) Posterior model probabilities via path-based pairwise priors. Statistica Neerlandica, 59:3-15.

Ozone data

```
Call:
lm(formula = y ~ ., data = ozone)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9981  -2.1254   0.0485   1.9645  12.4576

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.334e+03  3.071e+03  -1.411   0.1604
x4           1.593e+00  1.135e+00   1.403   0.1628
x5           3.868e+01  2.407e+01   1.607   0.1102
x6          -1.970e+00  3.167e+00  -0.622   0.5349
x7          -1.012e+01  8.183e+00  -1.237   0.2181
x8          -7.970e-03  2.614e-02  -0.305   0.7609
x9           2.102e+00  1.984e+00   1.060   0.2912
x10          9.778e-02  6.097e-01   0.160   0.8728
x4.x4       -1.462e-04  1.050e-04  -1.393   0.1658
x4.x5       -6.839e-03  4.403e-03  -1.553   0.1226
x4.x6        3.788e-04  5.700e-04   0.665   0.5074
x4.x7        1.814e-03  1.497e-03   1.212   0.2275
x4.x8        1.959e-06  4.761e-06   0.411   0.6814
x4.x9       -3.761e-04  3.665e-04  -1.026   0.3066
x4.x10      -2.138e-05  1.117e-04  -0.191   0.8485
x5.x5       -1.932e-01  8.951e-02  -2.158   0.0326 *
x5.x6       -1.271e-03  1.336e-02  -0.095   0.9243
x5.x7        3.698e-02  3.475e-02   1.064   0.2890
x5.x8        3.045e-07  1.522e-04   0.002   0.9984
x5.x9        2.732e-03  8.461e-03   0.323   0.7473
x5.x10       1.067e-03  2.577e-03   0.414   0.6793
x6.x6       -2.432e-03  1.804e-03  -1.348   0.1798
x6.x7        2.704e-03  4.483e-03   0.603   0.5473
x6.x8       -2.391e-05  1.988e-05  -1.203   0.2310
x6.x9        1.149e-03  1.474e-03   0.779   0.4371
x6.x10       2.705e-04  4.132e-04   0.655   0.5138
x7.x7       -2.334e-03  5.259e-03  -0.444   0.6579
x7.x8       -2.929e-05  3.916e-05  -0.748   0.4557
x7.x9        1.087e-04  2.830e-03   0.038   0.9694
x7.x10      -4.100e-04  1.010e-03  -0.406   0.6853
x8.x8       -1.908e-07  2.167e-07  -0.880   0.3801
x8.x9        5.260e-06  1.224e-05   0.430   0.6679
x8.x10       1.714e-06  3.724e-06   0.460   0.6461
x9.x9       -9.608e-04  4.583e-04  -2.096   0.0378 *
x9.x10      -2.062e-04  2.726e-04  -0.756   0.4507
x10.x10      4.983e-05  7.584e-05   0.657   0.5122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 4.137 on 142 degrees of freedom
Multiple R-squared:  0.7962,    Adjusted R-squared:  0.746
F-statistic: 15.85 on 35 and 142 DF,  p-value: < 2.2e-16
```

Ozone data

```
> stepAIC(lmfit,k=2,trace=FALSE)#AIC

Call:
lm(formula = y ~ x4 + x5 + x6 + x7 + x9 + x4.x4 + x4.x5 + x4.x6 +
    x4.x7 + x4.x9 + x5.x5 + x5.x7 + x6.x6 + x6.x8 + x6.x9 + x7.x10 +
    x9.x9 + x10.x10, data = ozone)

Coefficients:
(Intercept)           x4           x5           x6           x7
 -3.588e+03    1.323e+00    4.122e+01   -3.934e+00   -7.727e+00
         x9         x4.x4        x4.x5        x4.x6        x4.x7
  2.087e+00   -1.217e-04   -7.283e-03    7.424e-04    1.370e-03
      x4.x9        x5.x5        x5.x7        x6.x6        x6.x8
 -3.771e-04   -1.742e-01    3.651e-02   -2.243e-03   -1.039e-05
      x6.x9        x7.x10       x9.x9       x10.x10
  1.694e-03   -5.924e-04   -9.081e-04    7.793e-05

> stepAIC(lmfit,k=3,trace=FALSE)#BIC

Call:
lm(formula = y ~ x4 + x5 + x6 + x7 + x9 + x4.x4 + x4.x5 + x4.x6 +
    x4.x7 + x4.x9 + x5.x5 + x6.x6 + x6.x8 + x6.x9 + x7.x10 +
    x9.x9 + x10.x10, data = ozone)

Coefficients:
(Intercept)           x4           x5           x6           x7
 -2.707e+03    1.023e+00    1.708e+01   -3.966e+00   -6.528e+00
         x9         x4.x4        x4.x5        x4.x6        x4.x7
  2.152e+00   -9.660e-05   -2.799e-03    7.523e-04    1.191e-03
      x4.x9        x5.x5        x6.x6        x6.x8        x6.x9
 -3.910e-04   -1.153e-01   -2.446e-03   -1.101e-05    1.901e-03
     x7.x10       x9.x9       x10.x10
 -5.450e-04   -9.225e-04    6.887e-05

> leaps(ozone[,-1],ozone[,1]) #Forward stepwise selection
Error in leaps(ozone[, -1], ozone[, 1]) :
  leaps does not allow more than 31 variables; use regsubsets()
```

Subset selections using AIC (top) and BIC (middle).
Forward stepwise selection fails a data with more than 31 variables.

Ridge regression

Ridge regression is like least squares but shrinks the estimated coefficients towards zero. Given a response vector $Y \in \mathbb{R}^n$ and a predictor matrix $X \in \mathbb{R}^{n \times p}$, the regression coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ are defined as

$$\hat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} (y_i - x_i \boldsymbol{\beta})^2}_{\text{Loss}} + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{\text{Penalty}}$$

Turning parameter $\lambda \geq 0$ controls the strength of the penalty term.

- $\hat{\boldsymbol{\beta}}^{ridge}$ is the least square estimate when $\lambda = 0$.
- $\hat{\boldsymbol{\beta}}^{ridge} = 0$ when $\lambda = \infty$.
- Turning λ between a linear model (Y on X) and shrinking the coefficients.

Ridge regression

The ridge regression estimator is

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{ridge} &= \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \\ &= (X^T X + \lambda I_p)^{-1} X^T Y\end{aligned}$$

Proof: The first derivative is

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\|Y - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2) = 2X^T X\boldsymbol{\beta} - 2X^T Y + 2\lambda\boldsymbol{\beta}$$

and is 0 when $\boldsymbol{\beta} = (X^T X + \lambda I_p)^{-1} X^T Y$.

Ridge regression

An intercept β_0 is usually unpenalized and the ridge regression with intercept $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ is

$$\hat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - x_i\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

If Y and X are centered, $\hat{\beta}_0 = 0$ and no need to include an intercept, β_0.

The penalty term $||\boldsymbol{\beta}||_2^2$ is unfair to predict if all covariates are not on the same scale.

Therefore, the columns of X are standardized; Standardized design matrix $\tilde{X}$ such that $\widetilde{X} = 0$ and $\mathrm{Var}(\widetilde{X}) = 1$. Often both Y and X are standardized to perform the ridge regression.

Implementation using R

To fit the ridge regression using R:

```
library(glmnet)
fit = glmnet(x, y, data, lambda, alpha)
```

glmnet-function fits the elastic net. It is equivalent to the ridge regression when $\alpha = 0$.

See the manual for details;
https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

Ozone data

Coefficients and mean squared error for the ridge regression model with $\lambda = 1$.

```
> ridge.lam1=glmnet(ozone[,-1],ozone[,1],alpha=0,lambda=1,standardize=TRUE)
> ridge.lam1$beta
35 x 1 sparse Matrix of class "dgCMatrix"
                      s0
x4          3.291037e-03
x5          6.340630e-02
x6          2.058435e-02
x7          6.849918e-02
x8          5.291729e-05
x9          1.138298e-02
x10        -1.606820e-03
x4.x4       3.662978e-07
x4.x5       4.865389e-06
x4.x6       4.648775e-06
x4.x7       1.287906e-05
x4.x8      -1.685758e-09
x4.x9       1.860875e-06
x4.x10     -5.980880e-07
x5.x5      -1.729188e-02
x5.x6      -5.003315e-04
x5.x7      -8.655902e-05
x5.x8      -5.704469e-07
x5.x9       1.866369e-03
x5.x10      8.950660e-04
x6.x6       5.352922e-06
x6.x7       7.983332e-04
x6.x8      -8.912942e-06
x6.x9       2.710850e-05
x6.x10     -2.164049e-05
x7.x7       9.813809e-04
x7.x8      -9.979639e-06
x7.x9      -9.105093e-05
x7.x10     -3.040522e-04
x8.x8      -4.095447e-09
x8.x9      -1.467229e-06
x8.x10      2.170357e-06
x9.x9      -5.741249e-04
x9.x10      6.754323e-06
x10.x10     8.930635e-06
```

Ozone data

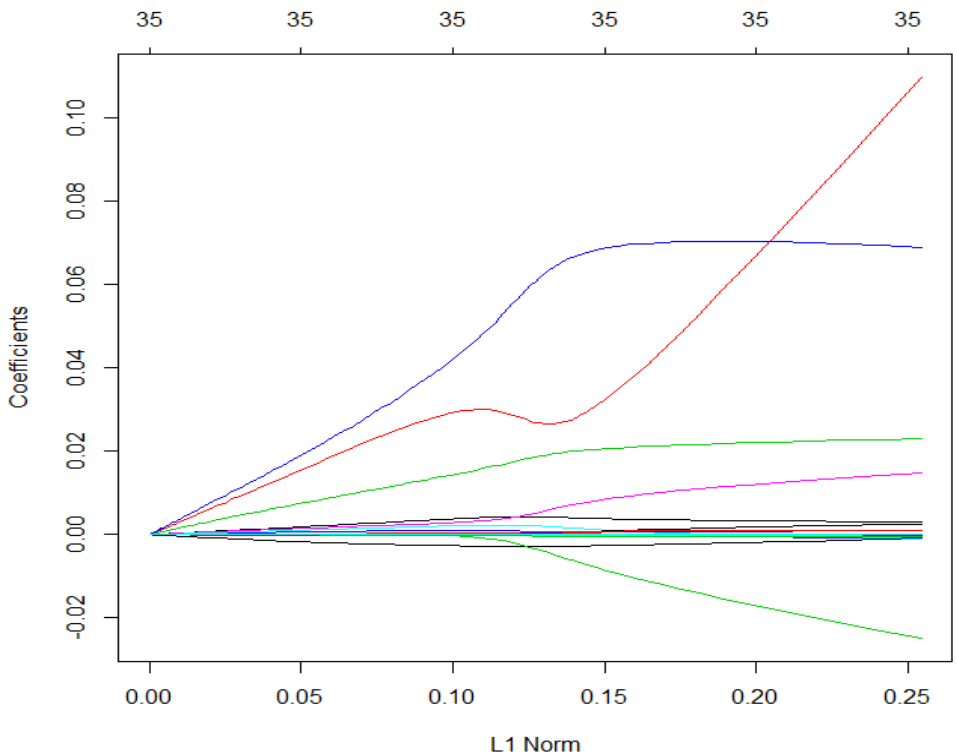
glmnet also estimates ridge regression models using different λ values. By details it computes for 100 λ values.

```
> glmnet(ozone[,-1],ozone[,1],alpha=0,standardize=TRUE,nlambda=10)

Call:  glmnet(x = ozone[, -1], y = ozone[, 1], alpha = 0, nlambda = 10,      standardize = TRUE)

      Df      %Dev    Lambda
 [1,] 35 1.416e-35 6523.0000
 [2,] 35 4.633e-02 2344.0000
 [3,] 35 1.176e-01  842.5000
 [4,] 35 2.614e-01  302.8000
 [5,] 35 4.588e-01  108.8000
 [6,] 35 6.135e-01   39.1000
 [7,] 35 6.872e-01   14.0500
 [8,] 35 7.199e-01    5.0500
 [9,] 35 7.413e-01    1.8150
[10,] 35 7.584e-01    0.6523
```

Ridge regression models using 10 λ values.

Ozone data

Ridge regression models using 100 λ-values.



Each curve corresponds to a variable and shows the path of its coefficient against l_1 norm ($\sum_j |\beta_j|$) with λ. The axis above is the number of nonzero coefficients at λ.

Ridge regression

The amount of bias of $\hat{\boldsymbol{\beta}}^{ridge}$ is

$$\mathbb{E}[\hat{\boldsymbol{\beta}}^{ridge} - \mathbb{E}[\boldsymbol{\beta}]] = ((X^T X + \lambda I)^{-1} X^T X - I) - \boldsymbol{\beta}$$

and the covariate matrix is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^{ridge}) = (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)\,.$$

There exists a value of λ for which the mean squared error of the ridge estimator is less than that of the least squares estimator.

Unfortunately, the appropriate value of λ depends on knowing the true regression coefficients (which are being estimated) and an analytic solution has not been found that guarantees the optimality of the ridge solution.

Least Absolute Selection and Shrinkage Operator (LASSO)

The LASSO regression coefficient $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ is found by

$$\hat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \underbrace{\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i\boldsymbol{\beta})^2}_{\text{Loss}} + \lambda \underbrace{\sum_{j=1}^{p}|\beta_j|}_{\text{Penalty}}$$

- Tuning parameter λ controls the strength of the penalty.
- $\hat{\boldsymbol{\beta}}^{lasso}$ is the linear regression (least square estimate) when $\lambda = 0$.
- $\hat{\boldsymbol{\beta}}^{lasso} = 0$ (null model) when $\lambda = \infty$.

λ is chosen to balance between a linear model with X (all covariates) and shrieked coefficients.

Least Absolute Selection and Shrinkage Operator (LASSO)

When including an intercept term in the model, we usually leave this coefficient unpenalized, just as we do with ridge regression. Hence the lasso problem with intercept $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ is

$$\hat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - x_i\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

If the columns of X are centered, the intercept estimate turns out to be $\hat{\beta}_0 = \bar{Y}$ (sample mean of Y). Therefore we typically center Y , X and do not include an intercept.

The penalty term $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ is not fair to predict if the predictor variables $x_1, ..., x_p$ are not on the same scale. It is advised to scale the columns of X then the lasso problem.

Least Absolute Selection and Shrinkage Operator (LASSO)

There is no explicit formulas for the bias and variance of the lasso estimate (e.g., when the true model is linear). However, we may nd some relationships.

$$\hat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - x_i\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- The bias increases as λ (amount of shrinkage) increases.
- The variance decreases as λ (amount of shrinkage) increases.

In terms of prediction error (or mean squared error), the lasso performs comparably to ridge regression.

Implementation using R

To fit the lasso regression using R:

```
library(glmnet)
fit = glmnet(x, y, data, lambda, alpha)
```

glmnet-function fits the elastic net. It is equivalent to the lasso regression when $\alpha = 1$.

See the manual for details;
https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

Ozone data

Coefficients and mean squared error for the lasso regression model with $\lambda = 1$.

```
> lasso1=glmnet(ozone[,-1],ozone[,1],alpha=1,lambda=1)
> lasso1$df #no of included covariates
[1] 6
> lasso1$beta
35 x 1 sparse Matrix of class "dgCMatrix"
                    s0
x4          .
x5          .
x6          .
x7          .
x8          .
x9          .
x10         .
x4.x4       .
x4.x5       .
x4.x6       .
x4.x7       .
x4.x8    -0.00187215
x4.x9       .
x4.x10      .
x5.x5       .
x5.x6       .
x5.x7       .
x5.x8       .
x5.x9       .
x5.x10      .
x6.x6       .
x6.x7       2.07592367
x6.x8       .
x6.x9       .
x6.x10      .
x7.x7       3.70240152
x7.x8    -0.78680091
x7.x9       .
x7.x10   -0.17795539
x8.x8       .
x8.x9       .
x8.x10      .
x9.x9    -0.05539032
x9.x10      .
x10.x10     .
> lasso1$df #no of included covariates
[1] 6
> pred.lasso1=predict(lasso1,ozone[,-1],s=1,type="response") #fitted values with lam=1
> mean((pred.lasso1-ozone[,1])^2) #MSE
[1] 20.21349
```

Ozone data

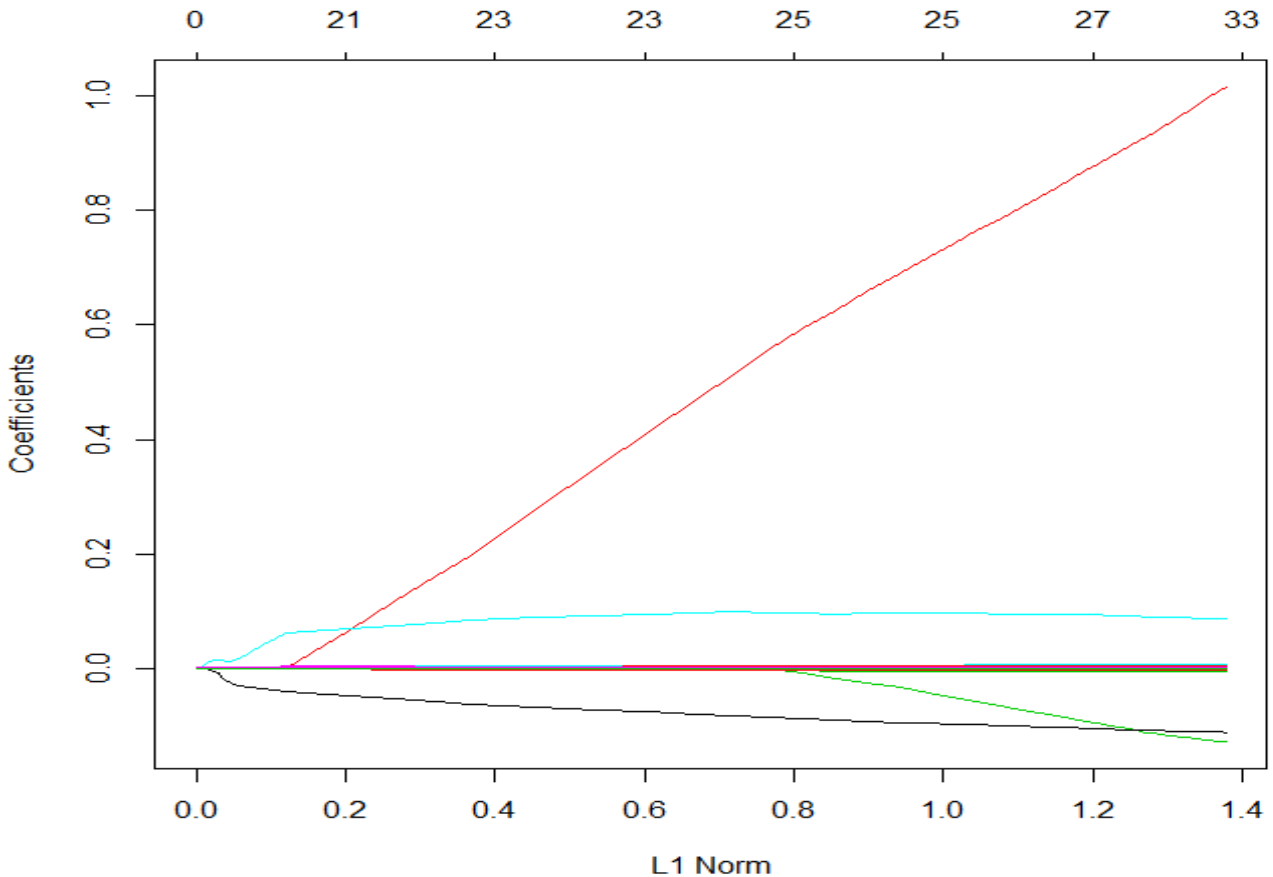
glmnet also estimates lasso regression models using different λ-values. By default $\lambda = 100$.

```
> glmnet(ozone[,-1],ozone[,1],alpha=1,standardize=TRUE,nlambda=10)

Call:  glmnet(x = ozone[, -1], y = ozone[, 1], alpha = 1, nlambda = 10,      standardize = TRUE)

      Df   %Dev    Lambda
 [1,]  0 0.0000 6.5230000
 [2,]  3 0.5907 2.3440000
 [3,]  5 0.7100 0.8425000
 [4,]  6 0.7375 0.3028000
 [5,] 10 0.7537 0.1088000
 [6,] 19 0.7698 0.0391000
 [7,] 25 0.7854 0.0140500
 [8,] 30 0.7897 0.0050500
 [9,] 28 0.7903 0.0018150
[10,] 33 0.7904 0.0006523
```

LASSO regression models using 10 λ-values.

Ozone data

LASSO regression models using 100 λ values.



Each curve corresponds to a variable and shows the path of its coefficient against l_1 norm ($\sum_j |\beta_j|$) with λ. The axis above is the number of nonzero coefficients at λ.

Ridge - vs - Lasso

- The nature of the l_2 penalty in Ridge model hardly causes zero coefficients.
- The nature of the l_1 penalty in LASSO causes some coefficients to be shrunken to zero exactly.
- In LASSO as λ increases, more coefficients are set to zero (less variables are selected), and among the nonzero coefficients, more shrinkage is employed.
- A model with zero coefficients is easier to interpret.

Ridge - vs - Lasso

Let's consider the problems with a constraint, C;

$$\hat{\beta}^{ridge} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (y_i - x_i\beta)^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq C$$

$$\hat{\beta}^{lasso} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} (y_i - x_i\beta)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq C$$

In comparison, the usual linear regression estimate solves the unconstrained least squares problem; these estimates constrain the coefficient vector to lie in some geometric shape centered around the origin.

This generally reduces the variance because it keeps the estimate close to zero. But which shape we choose really matters!

Degrees of freedom

The degrees of freedom of an estimate describes its effective number of parameters.
Let $X \in \mathbb{R}^{n \times p}$ be a fixed matrix of predictors.

- For linear regression, $\hat{Y} = X\hat{\beta}^{linear}$, $df(\hat{Y} = p)$.
- For ridge regression, $\hat{Y} = X\hat{\beta}^{ridge}$,

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

  then $df(\hat{Y}) = trace(X(X^TX + \lambda I_p)^{-1}X^T)$.

- For lasso, $\hat{Y} = X\hat{\beta}^{lasso}$,

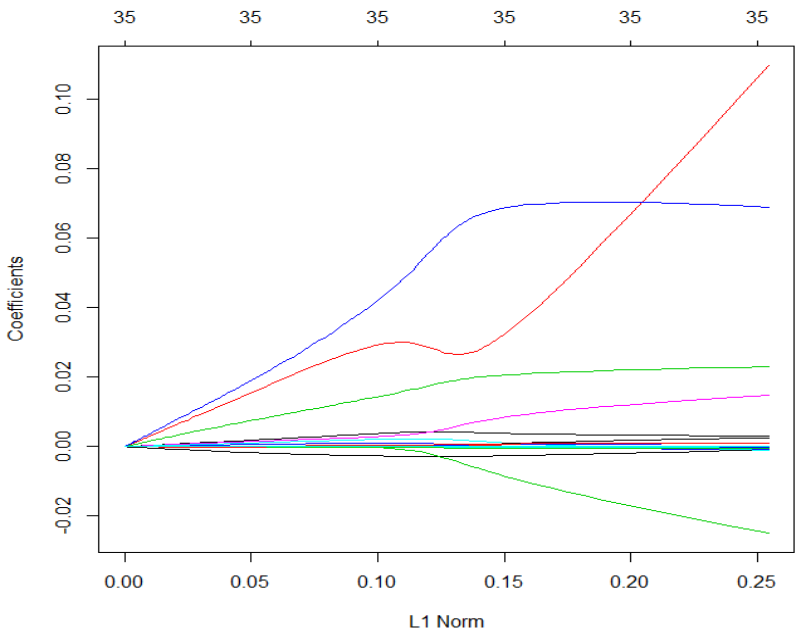
$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

  then $df(\hat{Y}) = \mathbb{E}[\text{number of nonzero coefficients in } \hat{\beta}^{lasso}]$

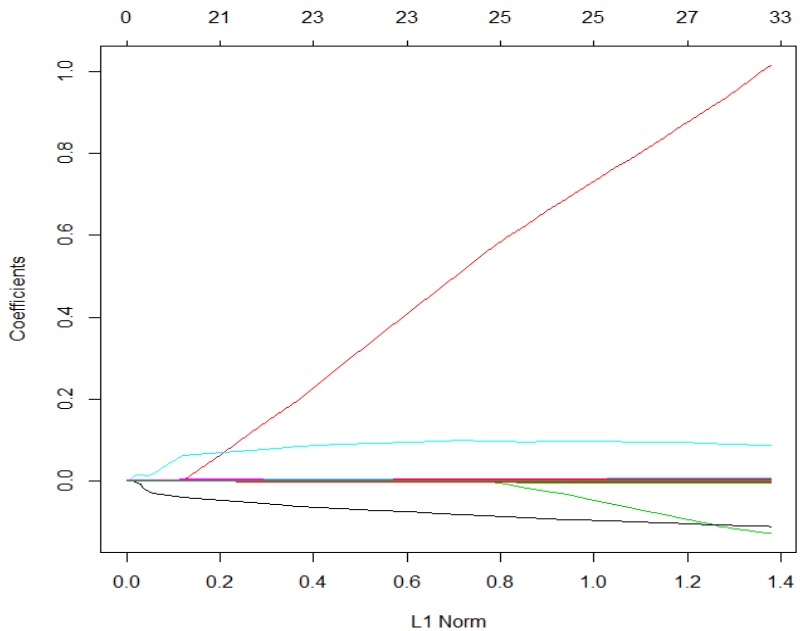
Ozone data

When $\lambda = 1$, the ridge regression model contains all 35 covariates and the lasso regression model contains only 6 covariates.

We can see the difference over λ-values.



Ridge mode



Lasso model

Penalized optimization

The penalty depends on the choice for λ. The question is what λ value is the best?

We choose λ maximizing the predictability using k-fold cross-validation. The mean squared error is often used for the loss function.

Penalized optimization

Let e_l be the mean squared error for the l-block based ridge/lasso regression model with λ. The prediction error of the k-fold c.v estimation is

$$CV(\lambda, k) = \frac{1}{k} \sum_{l=1}^{k} e_l \,.$$

Given m potential λ-values $(\lambda_1, ..., \lambda_m)$, the best $\hat{\lambda}$ minimizing the predictive error is chosen by

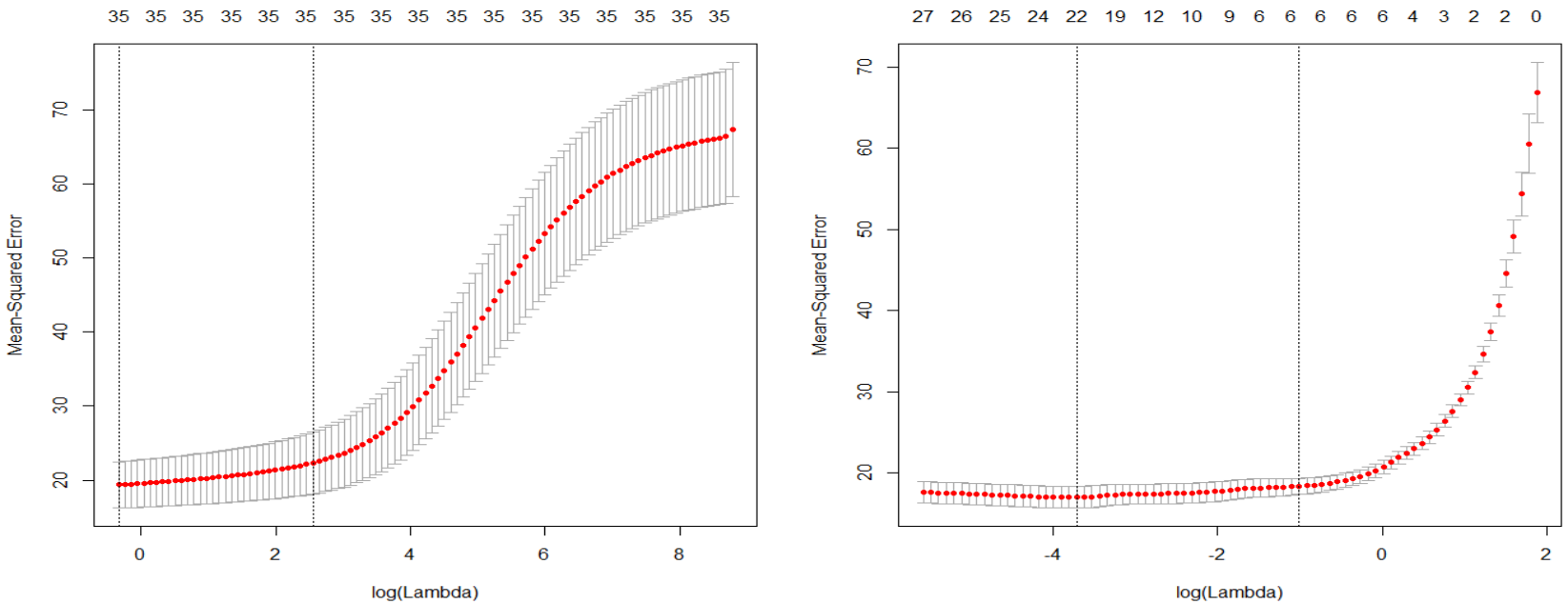
$$\hat{\lambda} = \min_{\lambda \in (\lambda_1, ..., \lambda_m)} CV(\lambda, k) \,.$$

If you prefer a parsimonious model, find $\hat{\lambda}$ and move λ in the direction of increasing λ until the cross-validation error curve is while 1sd of $CV(\lambda, k)$;

$$CV(\hat{\lambda}_{1se}) \leq CV(\hat{\lambda}) + \sigma_{CV(\hat{\lambda})}$$

Ozone data

MSE over λ values using 10-fold cross validation for (Left) ridge and (Right) lasso regression models.



Ridge model: $\hat{\lambda} = 0.7158862$ and $\hat{\lambda}_{1se} = 4.601769$ (35 covariates)
Lasso model: $\hat{\lambda} = 0.0203887$ (22 covariates) and $\hat{\lambda}_{1se} = 0.6372927$ (6 covariates)

Ozone data

Intercept & coefficient estimate;

```
> cbind(coef(cv.ridge, lambda=cv.ridge$lambda.1se),coef(cv.lasso, lambda=cv.lasso$lambda.1se))
36 x 2 sparse Matrix of class "dgCMatrix"
                      1             1
(Intercept) -3.756248e+01  5.598685e-01
x4           4.197537e-03  .
x5           2.653007e-02  .
x6           1.924239e-02  .
x7           6.347997e-02  .
x8          -1.164812e-04  .
x9           5.932593e-03  .
x10         -2.828357e-03  .
x4.x4        3.938394e-07  .
x4.x5        4.751217e-06  .
x4.x6        3.773747e-06  .
x4.x7        1.068102e-05  .
x4.x8       -2.281625e-08  .
x4.x9        1.004128e-06  .
x4.x10      -5.587089e-07  .
x5.x5       -4.416107e-03  .
x5.x6        4.917534e-04  .
x5.x7        1.890521e-03  .
x5.x8       -1.253801e-05  .
x5.x9        4.029207e-04  .
x5.x10       2.498876e-04  .
x6.x6        1.396378e-04  .
x6.x7        5.039026e-04  1.291920e-03
x6.x8       -4.509725e-06 -1.539454e-06
x6.x9        3.990339e-06  .
x6.x10      -3.367895e-05  .
x7.x7        6.314621e-04  2.184336e-03
x7.x8       -4.204451e-06 -8.218989e-06
x7.x9        6.185297e-05  .
x7.x10      -1.001466e-04 -1.024641e-04
x8.x8       -1.735190e-08  .
x8.x9       -1.572204e-06  .
x8.x10       5.910224e-07  .
x9.x9       -3.944730e-04 -1.995554e-04
x9.x10       2.259869e-05  .
x10.x10     -1.330737e-07  .
```

Ridge model: CV predictive error 18.62121 when $\hat{\lambda}$ and 20.097 when $\hat{\lambda}_{1se}$.
Lasso model: CV predictive error 17.77875 when $\hat{\lambda}$ and 20.44322 when $\hat{\lambda}_{1se}$.

More about LASSO

There are various approaches to improve the LASSO estimator.

- Bootstrapping Lasso Estimators
  - Use the modified versions of bootstrap to approximate the distribution of the Lasso estimator. [2]
- λ choice by the model selection criteria [3]
- When p is very large compared to n; $p \gg n$ [4]

---

[2] A. Chatterjee & S. N. Lahiri (2011) Bootstrapping Lasso Estimators, Journal of the American Statistical Association, 106:494, 608-625
M. Giurcanu & B. Presnell (2019) Bootstrapping LASSO-type estimators in regression models, Journal of Statistical Planning and Inference, 199, 114125

[3] H. Zou, T. Hastie, & R. Tibshirani (2007) ON THE "DEGREES OF FREEDOM" OF THE LASSO, The Annals of Statistics, Vol. 35, No. 5, 21732192

[4] E. Candes & T. Tao (2007) THE DANTZIG SELECTOR: STATISTICAL ESTIMATION WHEN p IS MUCH LARGER THAN n, The Annals of Statistics, Vol. 35, No. 6, 23132351

More about LASSO

The LASSO does not handle highly correlated variables very well; the coefficient paths tend to be erratic and can sometime show wild behaviour.

Example : There exists a pair of identical covariates. l_1 penalty is indifferent and the coefficients are not defined. l_2 will yield two equal valued coefficients.

One suggestions is a method between Ridge and Lasso.

Elastic Net

Compromise between the ridge and the lasso penalties and the estimate is defined as

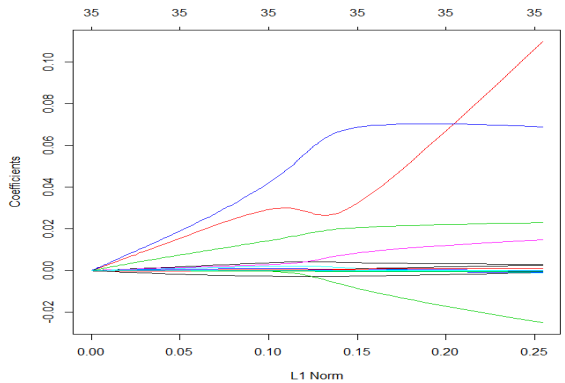
$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - x_i \boldsymbol{\beta})^2 + \lambda \left(\frac{1}{2}(1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right) \right\}$$

where $\alpha \in [0, 1]$.

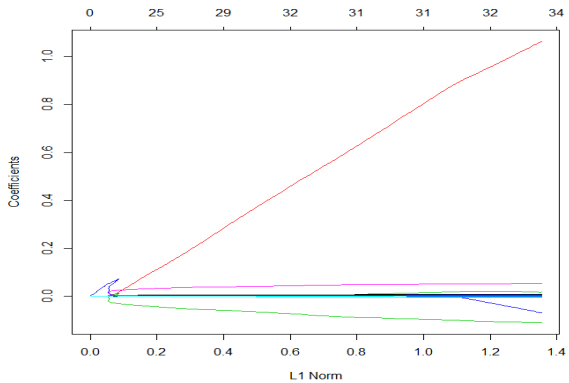
When $\alpha = 1$, it reduces to the l_1-norm (LASSO) and with $\alpha = 0$, it becomes to l_2-norm (ridge).

By adding some component of the ridge penalty to the l_1-penalty, the elastic net controls for within-group correlations.

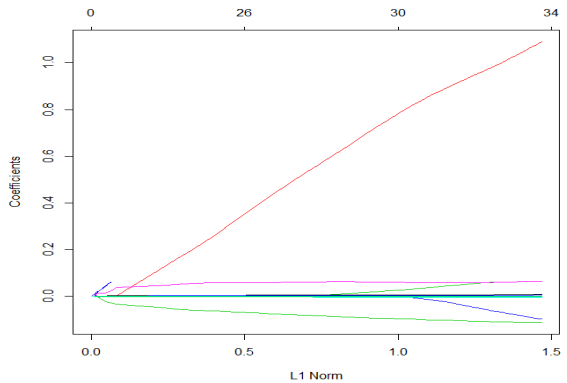
Ozone data



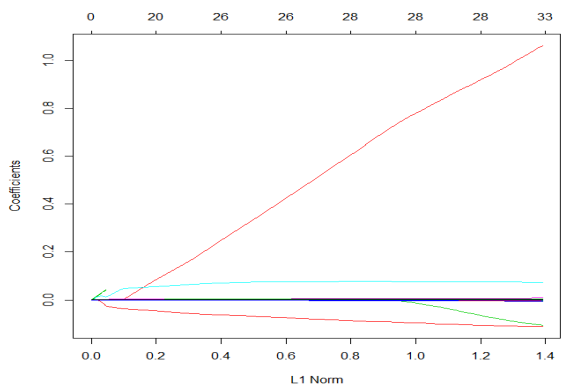
$\alpha = 0$ (ridge)



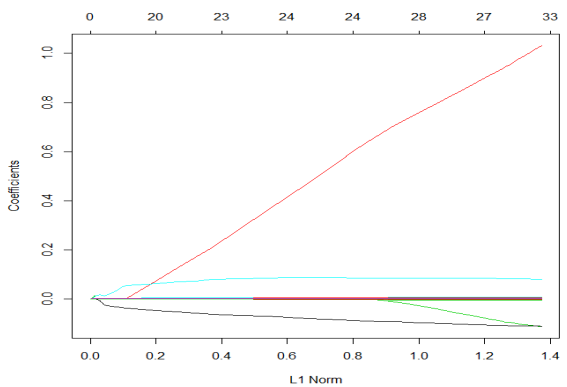
$\alpha = 0.2$



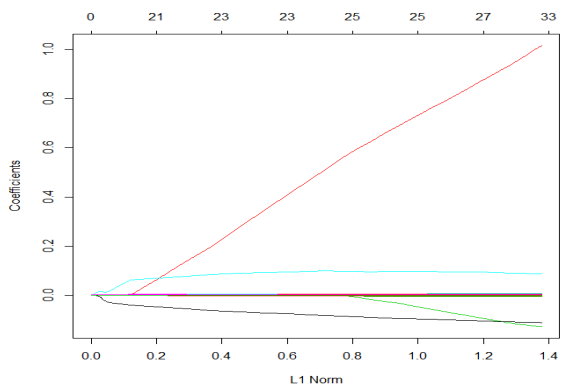
$\alpha = 0.4$



$\alpha = 0.6$

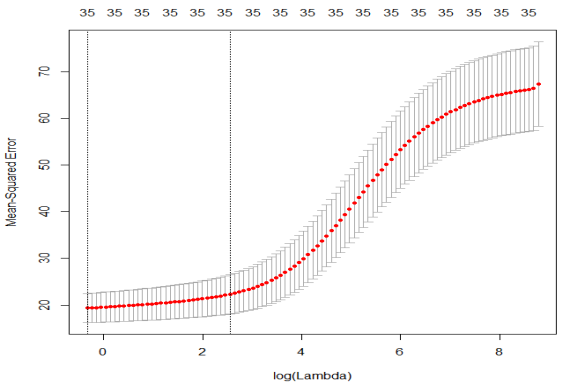


$\alpha = 0.8$

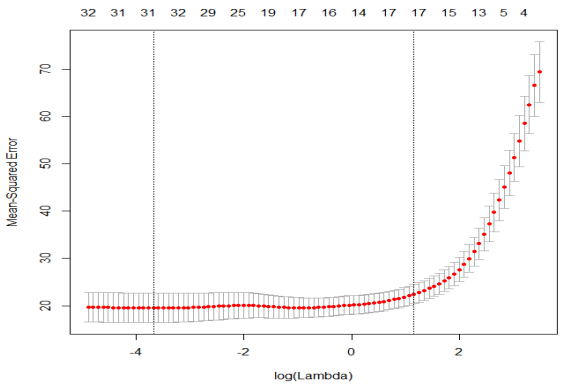


$\alpha = 1$ (lasso)

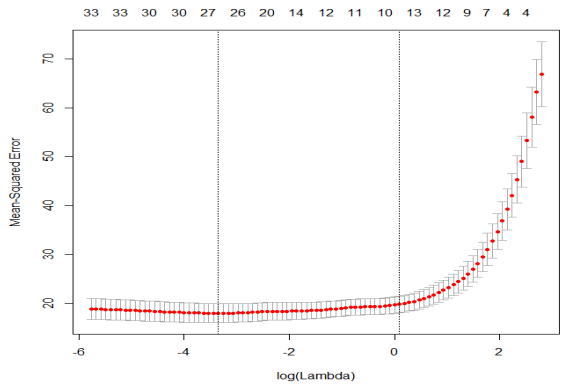
Ozone data



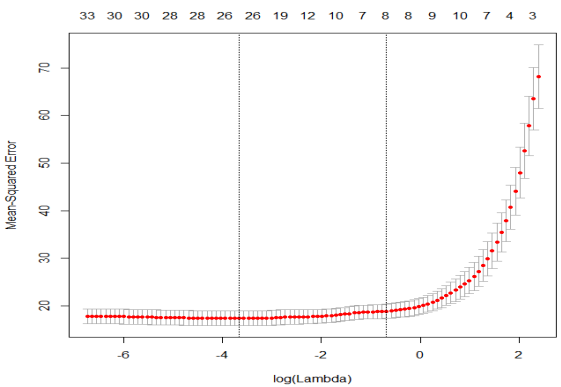
$\alpha = 0$ (ridge)



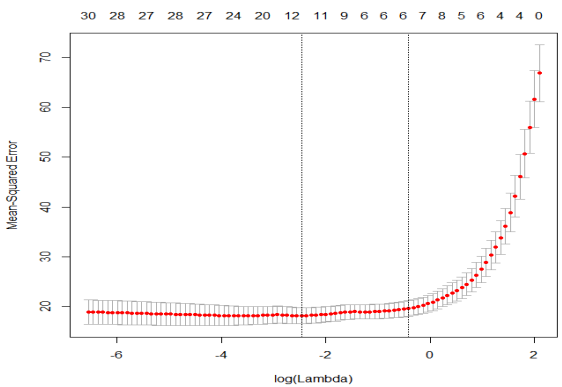
$\alpha = 0.2$



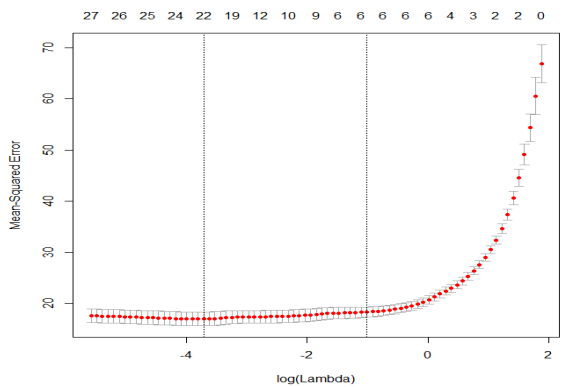
$\alpha = 0.4$



$\alpha = 0.6$



$\alpha = 0.8$



$\alpha = 1$ (lasso)

Ozone data

As α approaches to 1 from 0, the number of variables reduces to 22 by the min error criterion and 6 by the 1sd rule.

i.e., As $\alpha \to 1$, more variables have zero-coefficients.

As α approaches to 0, the number of variables approaches to 35.

i.e., As $\alpha \to 0$, more variables have non-zero coefficients.

Elastic Net

How to choose α?

One may think of the CV to find the optimal α. However, it is advised to be not a good idea.

It is advised to set α priory.

Categorical response variable

When a response variable is binary (0 or 1), a logistic regression model is useful to model the response variable.
i.e., In R, the GLM with family=binomial and link = ”logit”.

Variable selection method (ridge and lasso) is implemented by setting family distribution.

Graduate school data

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/dont admit, is a binary variable.

```
> head(binary)
admit gre  gpa rank
1     0 380 3.61    3
2     1 660 3.67    3
3     1 800 4.00    1
4     1 640 3.19    4
5     0 520 2.93    4
6     1 760 3.00    2
```

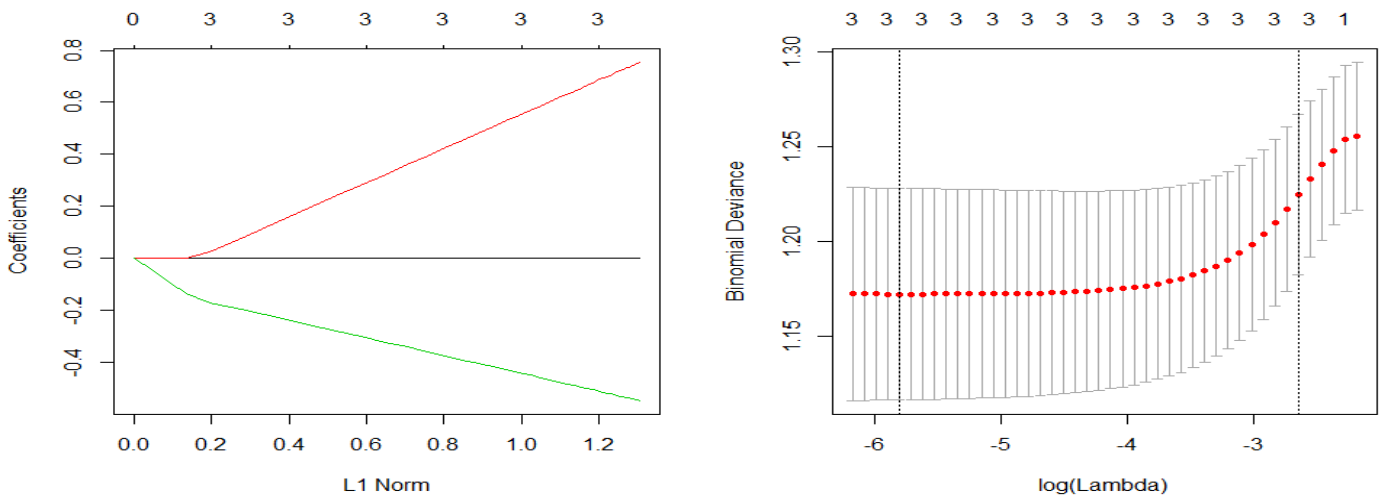
A logistic regression model is suitable to model the admission probability with GRE, GPA and Rank.

Graduate school data

```
binary.lasso=cv.glmnet(as.matrix(binary[,-1]),binary[,1],
family='binomial',alpha=1,standardize=TRUE)

#lambda minimizing the error
binary.lasso$lambda.1se

#coefficients when lambda.1se
coef(binary.lasso,s=binary.lasso$lambda.1se)

#prediction
predict(binary.lasso,as.matrix(binary[,-1]),lambda=binary.lasso$
```

Graduate school data

Coefficient estimates and CV error with λ values when l_1 penalty is applied (lasso).



cv error is 1.17 when $\hat{\lambda} = 0.0003$ and 1.22 when $\hat{\lambda}_{1sd} = 0.071$.

$$\log \frac{p(admi = 1)}{p(admi = 0)} = 0.734 + 0.0003gre + 0.085gpa - 0.201rank$$

Group LASSO

When covariates are categorical variables (group variable), how do we fit LASSO?

Penalty designed to have all coefficients within a group become nonzero (or zero).

Goal is to include or exclude this group of variables together.

Group LASSO

Given J groups, the vector $Z_j \in \mathbb{R}^{p_j}$ represents the covariates in group j. The goal is to predict $Y \in \mathbb{R}$ based on the collection of covariates $(Z_1, ..., Z_J)$.

i.e., Prediction of a linear regression is

$$\mathbb{E}[Y|Z, \theta] = \theta_0 + \sum_{j=1}^{J} \theta_j Z_j \,, \qquad \theta_j \in \mathbb{R}^{p_j}$$

i.e., If Y contains the effect of group j, $Z_k = 0$ for all $k \neq j$.

Group LASSO

Given n-samples $\{(y_i, z_{i1}, z_{i2}, ..., z_{iJ})\}_{i=1}^n$, the group lasso solves the convect problem

$$\underset{\theta}{\text{argmin}} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \theta_0 - \sum_{j=1}^{J} \theta_j z_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \|\theta_j\|_2$$

where $\|\theta_j\|_2$ is the Euclidean norm of the vector θ_j.

- Depending on λ, the entire vector θ_j will be zero or nonzero.
- When $p_j = 1$, $\|\theta_j\|_2 = \sqrt{\theta_j^2} = |\theta_j|$. Equivalent to the ordinary lasso.
- All J groups are equally penalized, a choice which leads larger groups to be more likely to be selected.

Group LASSO

```
library(grpreg)
fit <- grpreg(x, y, group, family= , penalty= )
cvfit <- cv.grpreg(x, y, group, family= , penalty="grLasso")
```

X	Design matrix without an intercept. grpreg standardizes the data and includes an intercept by default.
y	The response vector
group	A vector describing the grouping of the coefficients
penalty	The penalty to be applied to the model.
family	family distribution; ”gaussian”, ”binomial”, ”poisson”
lambda	λ value

Low infant birth weight

The Birthwt data contains 189 infant birth weights and low birth weight indicators.[5]

bwt	Birth weight in kilograms
low	Indicator of birth weight less than 2.5kg
X	Matrix of predictors
group	Vector describing how the columns of X are grouped Total 8 levels; age/lwt/race/smoke/ptl/ht/ui/ftv

[5] The data were collected at Baystate Medical Center, Springfield, Mass during 1986.

Low infant birth weight

X contains group-specific variables and the total is 16.

age1,age2,age3	Orthogonal polynomials of first, second, and third degree representing mothers age in years
lwt1,lwt2,lwt3	Orthogonal polynomials of first, second, and third degree representing mothers weight in pounds at last menstrual period
white,black	Indicator functions for mothers race; ”other” is reference group
smoke	Smoking status during pregnancy
ptl1,ptl2m	Indicator functions for one or for two or more previous premature labors, respectively. No previous premature labors is the reference category.
ht	History of hypertension
ui	Presence of uterine irritability
ftv1,ftv2,ftv3m	Indicator functions for one, for two, or for three or more physician visits during the first trimester, respectively. No visits is the reference category

Low infant birth weight

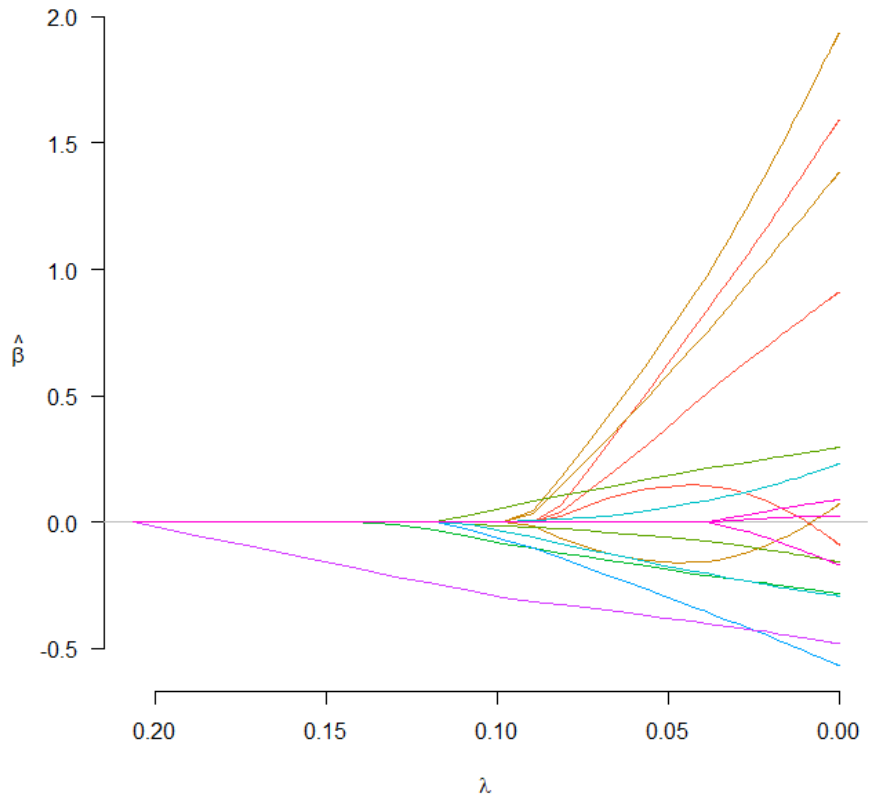
```
> str(Birthwt)
List of 4
 $ X    : num [1:189, 1:16] -0.0583 0.1344 -0.0446 -0.0308 -0.0721 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:16] "age1" "age2" "age3" "lwt1" ...
 $ bwt  : num [1:189] 2.52 2.55 2.56 2.59 2.6 ...
 $ low  : int [1:189] 0 0 0 0 0 0 0 0 0 0 ...
 $ group: Factor w/ 8 levels "age","lwt","race",..: 1 1 1 2 2 2 3 3 4 5 ...
> head(Birthwt$X)
            age1         age2         age3        lwt1        lwt2         lwt3
[1,] -0.05833434  0.011046300  0.029561818  0.12446282 -0.02133871 -0.130731102
[2,]  0.13436561  0.055245529 -0.096907046  0.06006722 -0.06922831 -0.033348413
[3,] -0.04457006 -0.009415469  0.045088774 -0.05918388  0.03746349  0.004618178
[4,] -0.03080577 -0.026243567  0.052489640 -0.05202881  0.02390664  0.019034579
[5,] -0.07209862  0.035141739  0.004821882 -0.05441384  0.02832410  0.014571538
[6,] -0.03080577 -0.026243567  0.052489640 -0.01386846 -0.03296942  0.049559472
     white black smoke ptl1 ptl2m ht ui ftv1 ftv2 ftv3m
[1,]     0     1     0    0     0  0  1    0    0     0
[2,]     0     0     0    0     0  0  0    0    0     1
[3,]     1     0     1    0     0  0  0    1    0     0
[4,]     1     0     1    0     0  0  1    0    1     0
[5,]     1     0     1    0     0  0  1    0    0     0
[6,]     0     0     0    0     0  0  0    0    0     0
> levels(Birthwt$group)
[1] "age"   "lwt"   "race"  "smoke" "ptl"   "ht"    "ui"    "ftv"
> head(Birthwt$bwt)
[1] 2.523 2.551 2.557 2.594 2.600 2.622
```

Low infant birth weight

We want to find the optimal model for the birth weight with 8 group variables and the penalty is l_2.

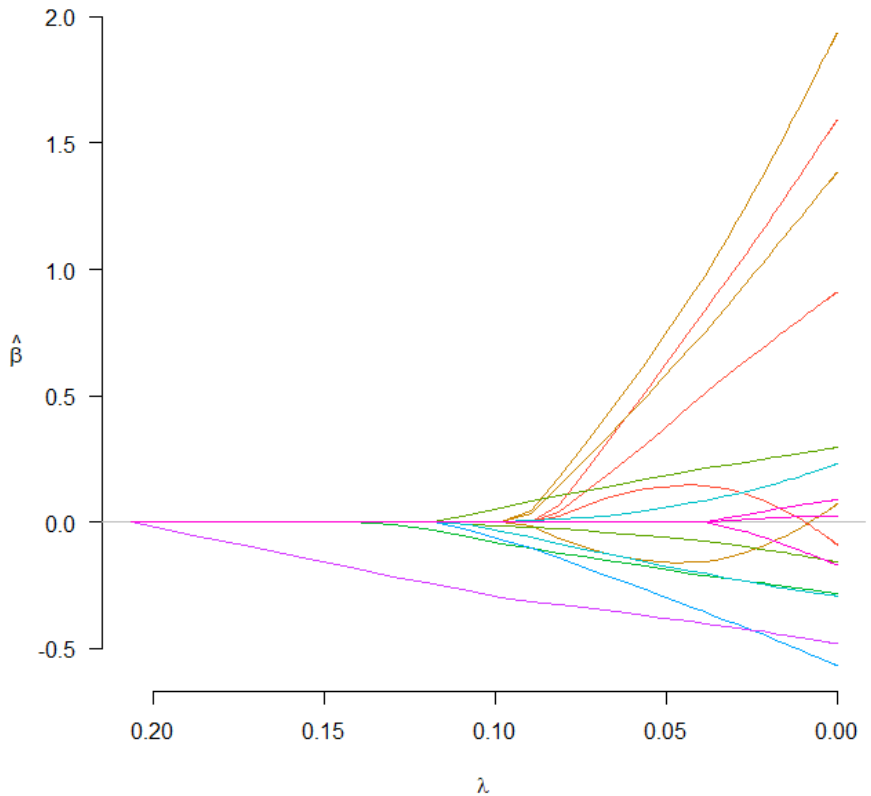
Low infant birth weight

```
fit <- grpreg(Birthwt$X, Birthwt$bwt, Birthwt$group)
plot(fit)
```



Low infant birth weight

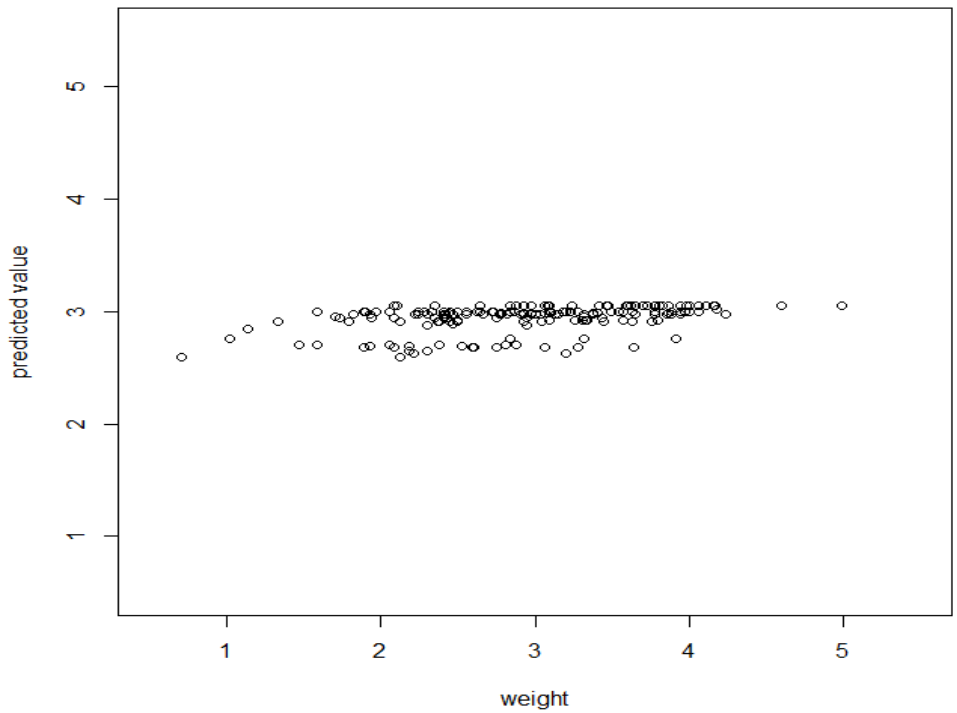
```
fit <- grpreg(Birthwt$X, Birthwt$bwt, Birthwt$group)
plot(fit)
```



Low infant birth weight

```
> #Fitted value
> predict(fit, Birthwt$X, type="response", lambda=0.1)[1:10]
 [1] 2.695672 3.002105 2.974998 2.682484 2.682484 3.002105 3.053843 3.002105
 [9] 2.974998 2.974998
> #Nonzero groups
> predict(fit, Birthwt$X, type="groups", lambda=0.1)
[1] "race"  "smoke" "ptl"   "ht"    "ui"
> #Coefficients
> predict(fit, Birthwt$X, type="coefficients", lambda=0.1)
                     0.1
(Intercept)  3.002104655
age1         0.000000000
age2         0.000000000
age3         0.000000000
lwt1         0.000000000
lwt2         0.000000000
lwt3         0.000000000
white        0.051738174
black       -0.013918906
smoke       -0.078844644
ptl1        -0.030111578
ptl2m        0.001479376
ht          -0.061421136
ui          -0.292514073
ftv1         0.000000000
ftv2         0.000000000
ftv3m        0.000000000
```

Low infant birth weight



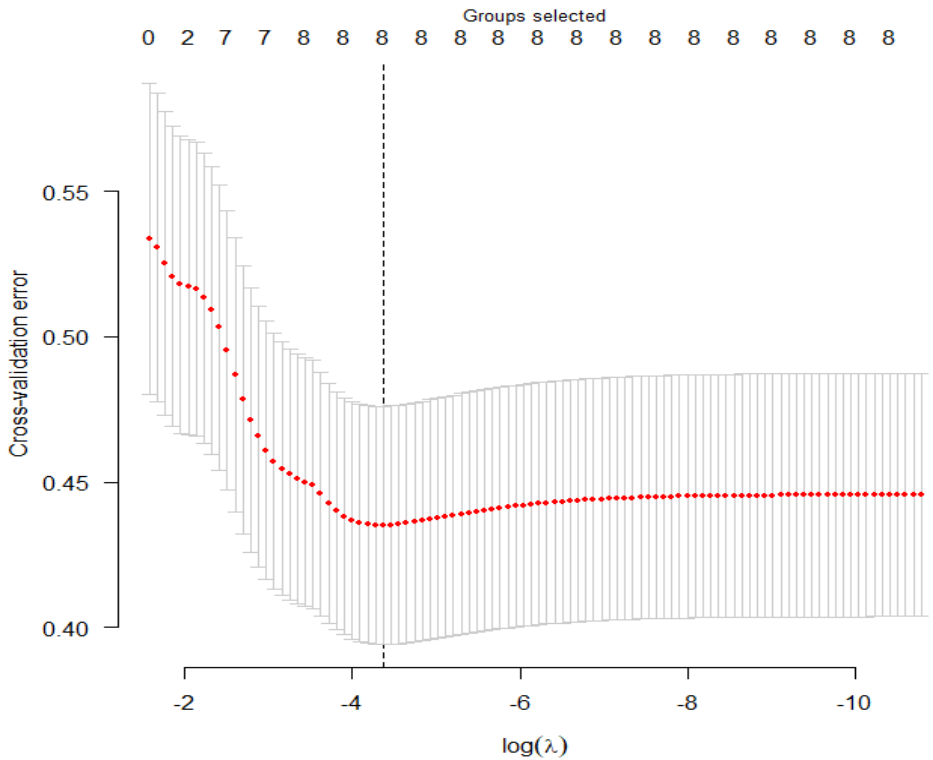
Very poor fit!
Let's find λ minimizing the error.

Low infant birth weight

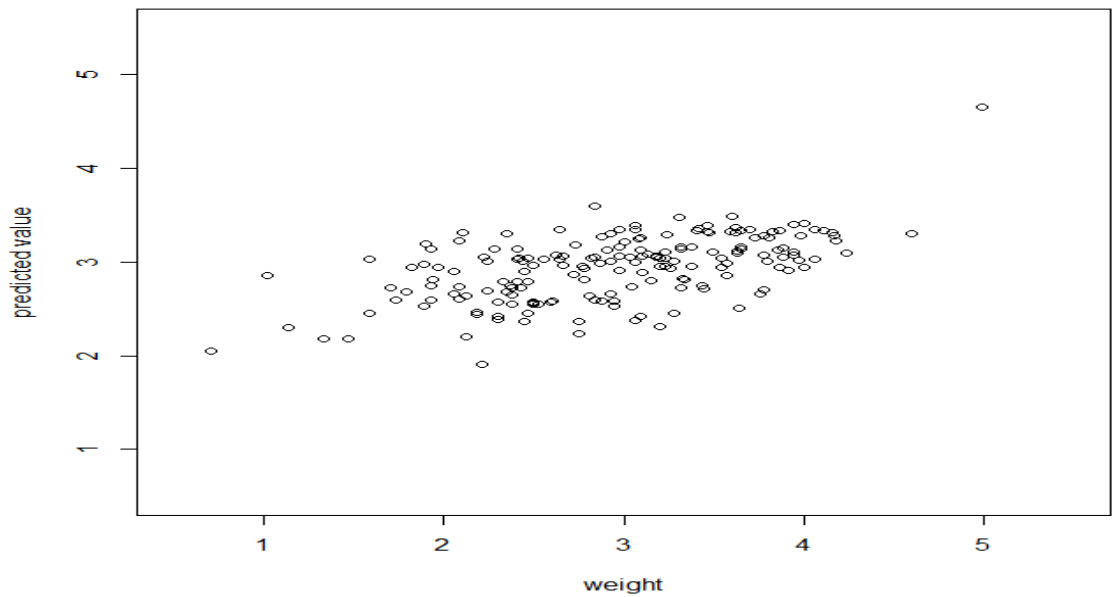
```
> cvfit <- cv.grpreg(Birthwt$X, Birthwt$bwt, Birthwt$group, penalty="grLasso")
> coef(cvfit)
(Intercept)        age1        age2        age3        lwt1        lwt2
3.03980161  0.06605057  1.22157782  0.72417976  1.45304814 -0.07667569
lwt3       white       black       smoke        ptl1       ptl2m
1.07987795  0.25511386 -0.11491470 -0.24778502 -0.25537334  0.15024454
ht          ui        ftv1        ftv2       ftv3m
-0.46412745 -0.44048368  0.05015767  0.01708853 -0.07517179
> cvfit$lambda.min
[1] 0.01838254
```

All groups are included in the model mininizing the error.

Low infant birth weight



Low infant birth weight



Sparse Group LASSO

With l_2 norm penalty, coefficients for any included group in a group-lasso fit, are nonzero.

Sometimes we would like sparsity both with respect to which groups are selected, and which coefficients are nonzero within a group.

$$\underset{\theta}{\text{argmin}} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} \theta_j z_{ij} \right)^2 + \lambda \sum_{j=1}^{J} \left((1-\alpha) \|\theta_j\|_2 + \alpha \|\theta_j\|_1 \right)$$

where $\alpha \in [0, 1]$.

When $\alpha = 0$, it becomes the ordinary group lasso.

When $\alpha = 1$, it becomes the ordinary lasso.

Reference

The Elements of Statistical Learning. Hastie, T., Tibshirani, R., Friedman, J. Spring Series in Statistics.

`http://web.stanford.edu/~hastie/ElemStatLearn/`

`https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html`