Statistical Learning. Linear methods for regression II

Jose Ameijeiras Alonso
Departamento de Estatística e Investigación Operativa (USC)

Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data

Multiple linear regression

- Suppose that we observe a quantitative response Y and preditor variables $X = (X_1, \ldots, X_p)$
- We write our model in general as

$$\boxed{Y = f(X) + \epsilon}$$

  where ϵ is a zero-mean error term that captures measurement errors and other discrepancies.
- The linear regression model assumes that

$$\boxed{f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}$$

Multiple linear regression

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon}$$

- Given a training sample $(y_1, x_{11}, \ldots, x_{1p}), \ldots, (y_n, x_{n1}, \ldots, x_{np})$, the ordinary least squares estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ solve the problem:

$$\boxed{\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2}$$

- The solution is given by:

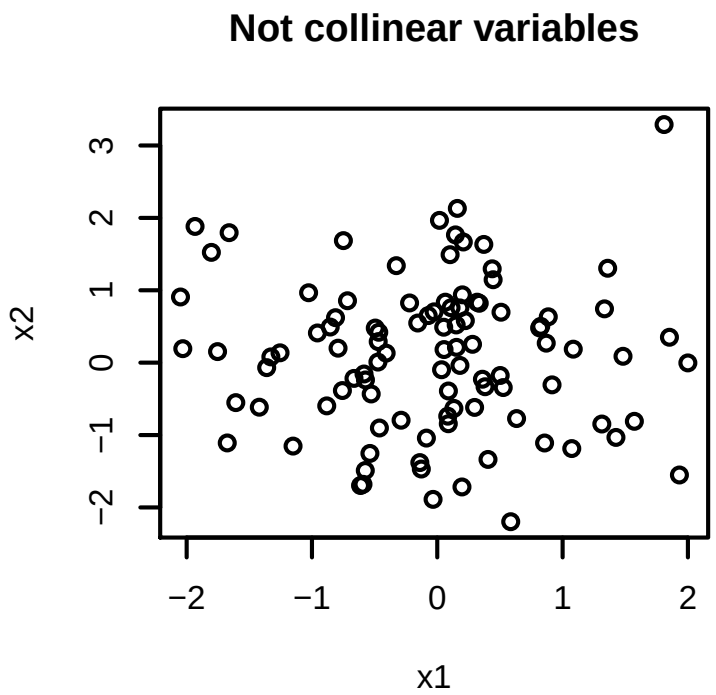
$$\boxed{\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t \mathbf{X}\right)^{-1} \mathbf{X}^t \mathbf{y}}$$

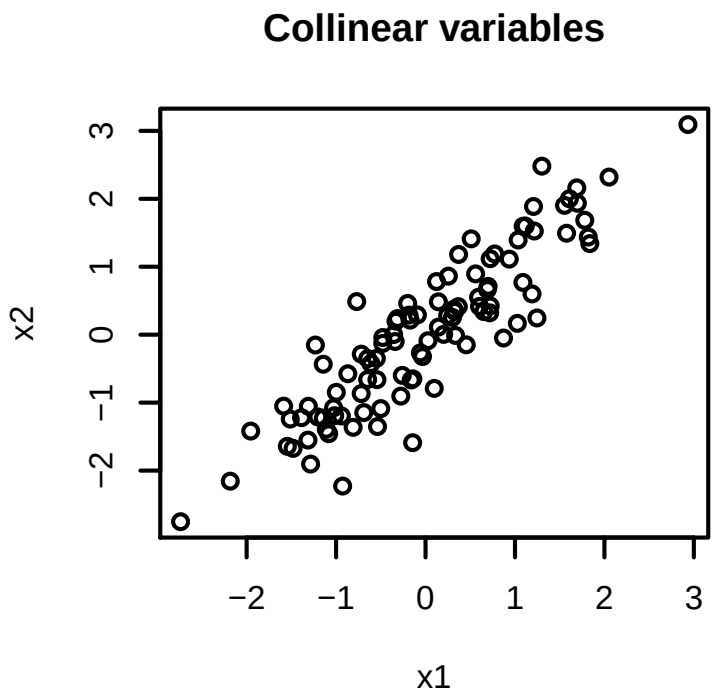
where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

Shortcomings of multiple linear regression. Multicollinearity

- Multicollinearity: is a problem that you can run into when you are fitting a regression model and two or more predictor variables are moderately or highly correlated (meaning that one can be linearly predicted from the others)





Shortcomings of multiple linear regression. Multicollinearity

Example: suppose we generate $n = 100$ observations from two variables X_1 and X_2 that are **not correlated**. The responses are drawn from the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

with $\beta_0 = 0$, $\beta_1 = 3$ and $\beta_2 = 5$.

```
> n <- 100
> beta <- c(3, 5)  # True parameters (intercept=0)
> x1 <- rnorm(n)
> x2 <- rnorm(n)
> y <- beta[1] * x1 + beta[2] * x2 + rnorm(n)
> mod <- lm(y ~ x1 + x2)
> coef(mod)

## (Intercept)          x1          x2
##   0.0273004   3.0163275   4.8602182
```

Shortcomings of multiple linear regression. Multicollinearity

Example: suppose we generate $n = 100$ observations from two variables X_1 and X_2 that are **highly correlated**. The responses are drawn from the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

with $\beta_0 = 0$, $\beta_1 = 3$ and $\beta_2 = 5$.

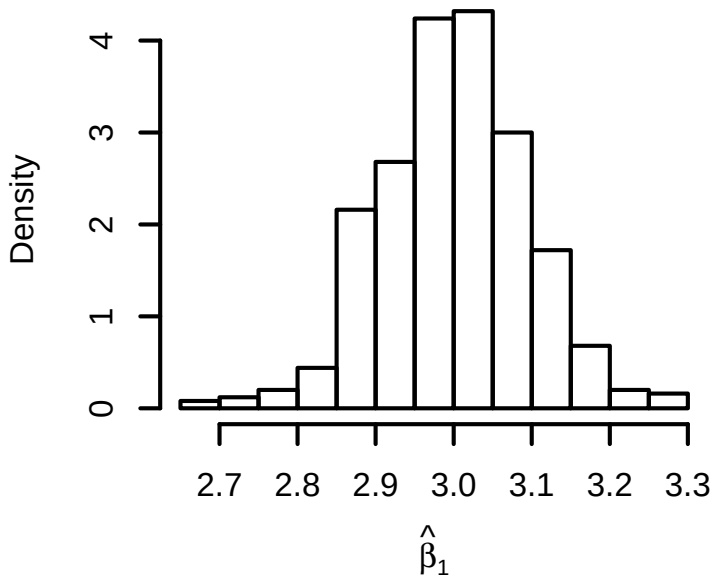
```
> n <- 100
> beta <- c(3, 5)  # True parameters (intercept=0)
> x1 <- rnorm(n)
> x2 <- rnorm(n, mean = x1, sd = 0.01)
> y <- beta[1] * x1 + beta[2] * x2 + rnorm(n)
> mod <- lm(y ~ x1 + x2)
> coef(mod)

## (Intercept)          x1          x2
##   -0.194826   21.125426  -13.140016
```

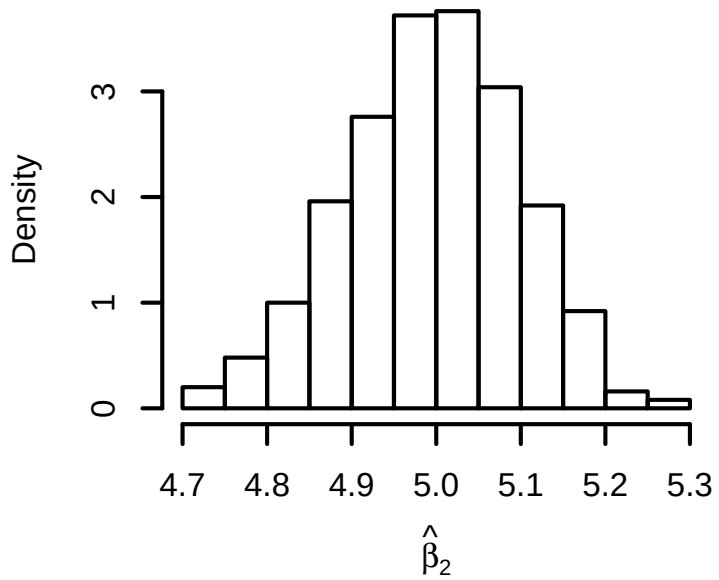
Shortcomings of multiple linear regression. Multicollinearity

Example: suppose we generate $n = 100$ observations from two variables X_1 and X_2 that are **not correlated**. The responses are drawn from the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

with $\beta_0 = 0$, $\beta_1 = 3$ and $\beta_2 = 5$. We repeat the experiment $B = 500$ times and represent a histogram of $\hat{\beta}_1$ and $\hat{\beta}_2$



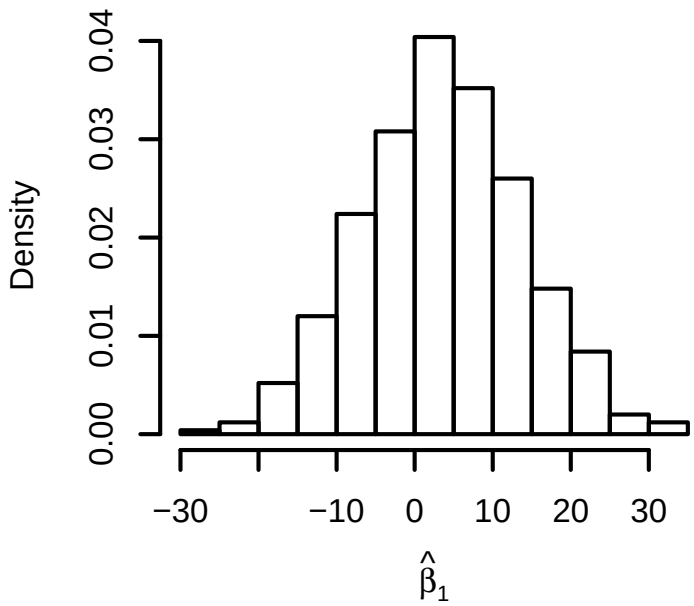


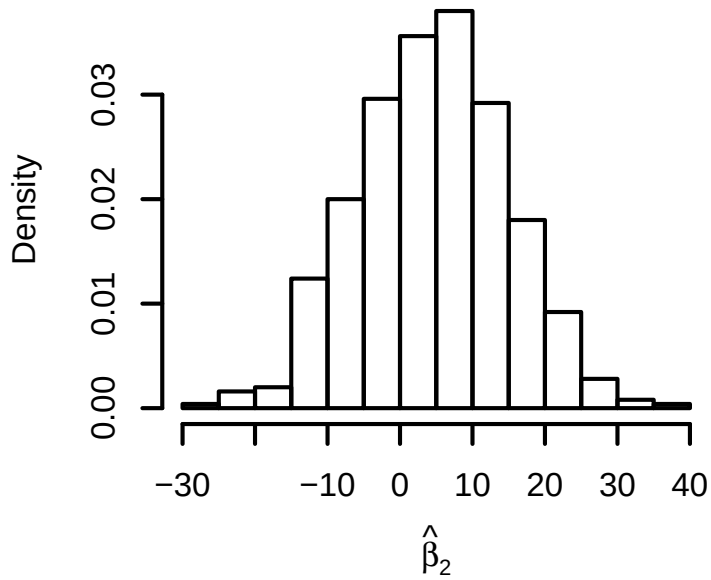
Shortcomings of multiple linear regression. Multicollinearity

Example: suppose we generate $n = 100$ observations from two variables X_1 and X_2 that are **highly correlated**. The responses are drawn from the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

with $\beta_0 = 0$, $\beta_1 = 3$ and $\beta_2 = 5$. We repeat the experiment $B = 500$ times and represent a histogram of $\hat{\beta_1}$ and $\hat{\beta_2}$





- The variance of multiple regression coefficients is inflated by the presence of correlated variables

Shortcomings of multiple linear regression. More features than observations

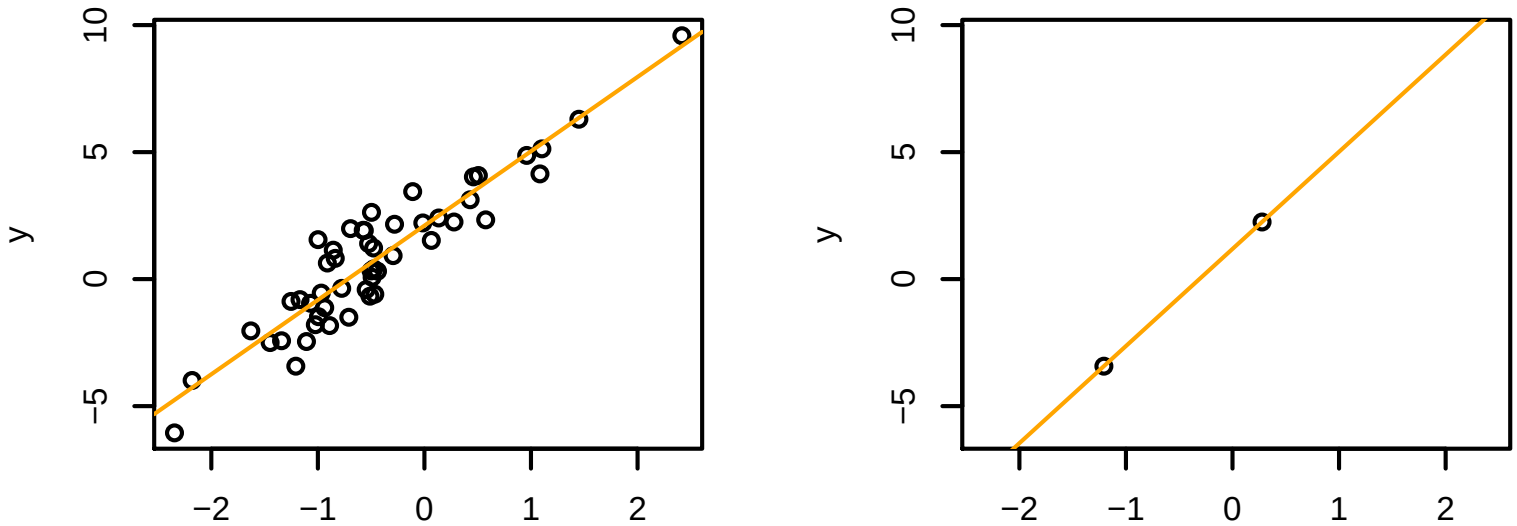
- More features than observations: classical approaches such as least squares linear regression are not appropriate when the number of features p is as large as, or larger than, the number of observations n
  - Most traditional statistical techniques for regression and classification are intended for the low-dimensional setting in which n, is much greater than p, the number of features
  - Data sets containing more features than observations are often referred to as high-dimensional

Shortcomings of multiple linear regression. More features than observations

Example: suppose we generate $n = 50$ (left) and $n = 2$ (right) observations from a variables X ($p = 1$). The responses are drawn from the model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

with $\beta_0 = 2$, $\beta_1 = 3$



- When $p > n$ or $p \approx n$, a simple least squares regression line is too flexible and overfits the data.

- It is possible to perfectly fit the training data in the high-dimensional setting but the resulting linear model will perform extremely poorly on an independent test set (poor prediction accuracy)

Alternatives to using least squares fit in the linear model

- The linear regression model assumes that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

- Before moving to non-linear models, we discuss some ways in which the simple linear model can be improved (better prediction accuracy and model interpretability)
  - Subset Selection. Methods for selecting a subset of the p predictors. We then fit a model using least squares on the reduced set of variables
    - Best subset selection, stepwise selection, AIC, BIC, adjusted R^2, cross-validation methods,...
  - Shrinkage (regularization). This approach involves fitting a model involving all p predictors. But the estimated coefficients are shrunken towards zero relative to the least squares estimates.
    - Ridge regression, Lasso,...
  - Dimension Reduction. Methods for projecting the p predictors into a lower-dimensional subspace. Then, the projections are used as predictors to fit a linear regression model by least squares.

Shrinkage methods

- Shrinkage methods fit a model containing all p predictors using a technique that shrinks the estimated coefficients towards zero
- Shrinking the coefficient estimates can significantly reduce their variance
- The two best-known regularization regression methods for linear regression are ridge regression and the Lasso

Ridge regression

- The linear regression model assumes that

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon}$$

  where ϵ is a zero-mean error term

- Given a training sample $(y_1, x_{11}, \ldots, x_{1p}), \ldots, (y_n, x_{n1}, \ldots, x_{np})$, the ordinary least squares estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ solve the problem

$$\boxed{\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2}$$

- The ridge regression estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ solve the problem

$$\boxed{\begin{array}{ll} \underset{\boldsymbol{\beta}}{\text{minimize}} & \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \\ \text{subject to} & \sum_{j=1}^{p} \beta_j^2 \leq s. \end{array}}$$

Ridge regression

- The linear regression model assumes that

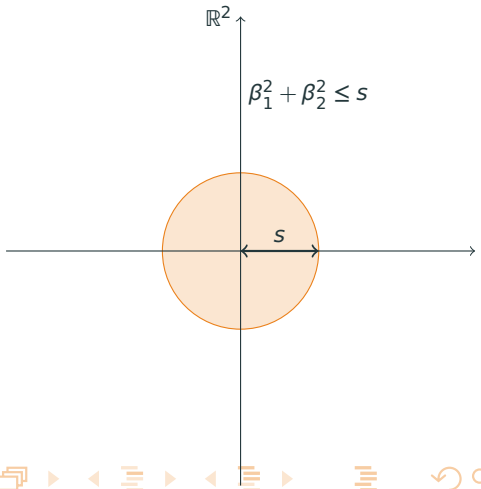
$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon}$$

  where ϵ is a zero-mean error term

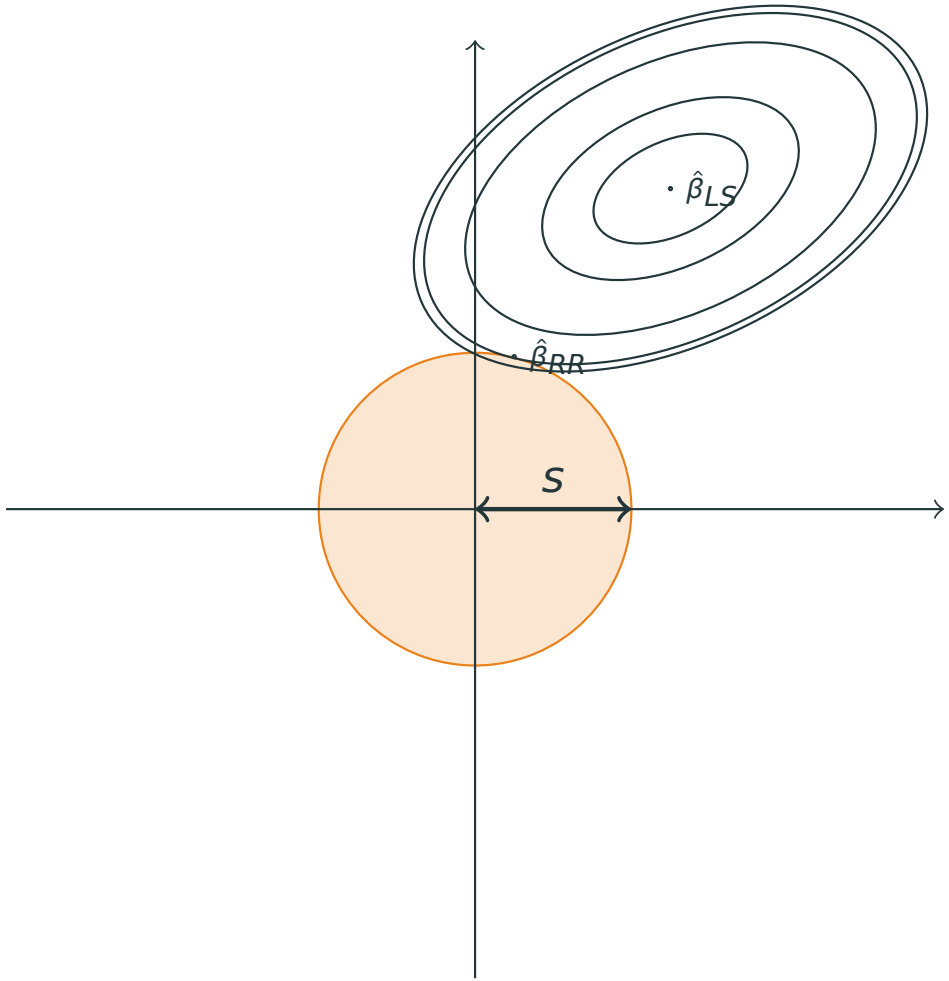
- The ridge regression estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ solve the problem

$$\boxed{\begin{aligned} &\underset{\boldsymbol{\beta}}{\text{minimize}} && \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \\ &\text{subject to} && \sum_{j=1}^{p} \beta_j^2 \leq s. \end{aligned}}$$

When $p = 2$, the ridge regression coefficient estimates have the smallest RSS out of all points that lie within the circle defined by $\beta_1^2 + \beta_2^2 \leq s$



Ridge regression



Ridge regression

- The linear regression model assumes that

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon}$$

  where ϵ is a zero-mean error term

- The ridge regression estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ solve the problem

$$\boxed{\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} \beta_j^2 \leq s. \end{aligned}}$$

- By the Lagrange multipliers method, the problem can be shown to be equivalent to:

$$\boxed{\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \beta_j^2.}$$

  (for every value of s there is some λ such that both optimization problems will give the same coefficient estimates)

Ridge regression

- The ridge regression estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ solve the problem

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

- Ridge regression seeks coefficient estimates that make the RSS small, but the second term (shrinkage penalty) has the effect of shrinking the estimates towards zero
- The parameter $\lambda \geq 0$ is a tuning parameter, to be determined separately
- The larger the value of λ, the greater the amount of shrinkage
- Ridge regression will produce a different set of coefficient for each value of λ

Ridge regression

- The ridge regression estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ solve the problem

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

- The shrinkage penalty is not applied to the intercept β_0
- The ridge solutions are not equivariant under scaling of the inputs. We should standardize the inputs before solving the optimization problem

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

Ridge regression

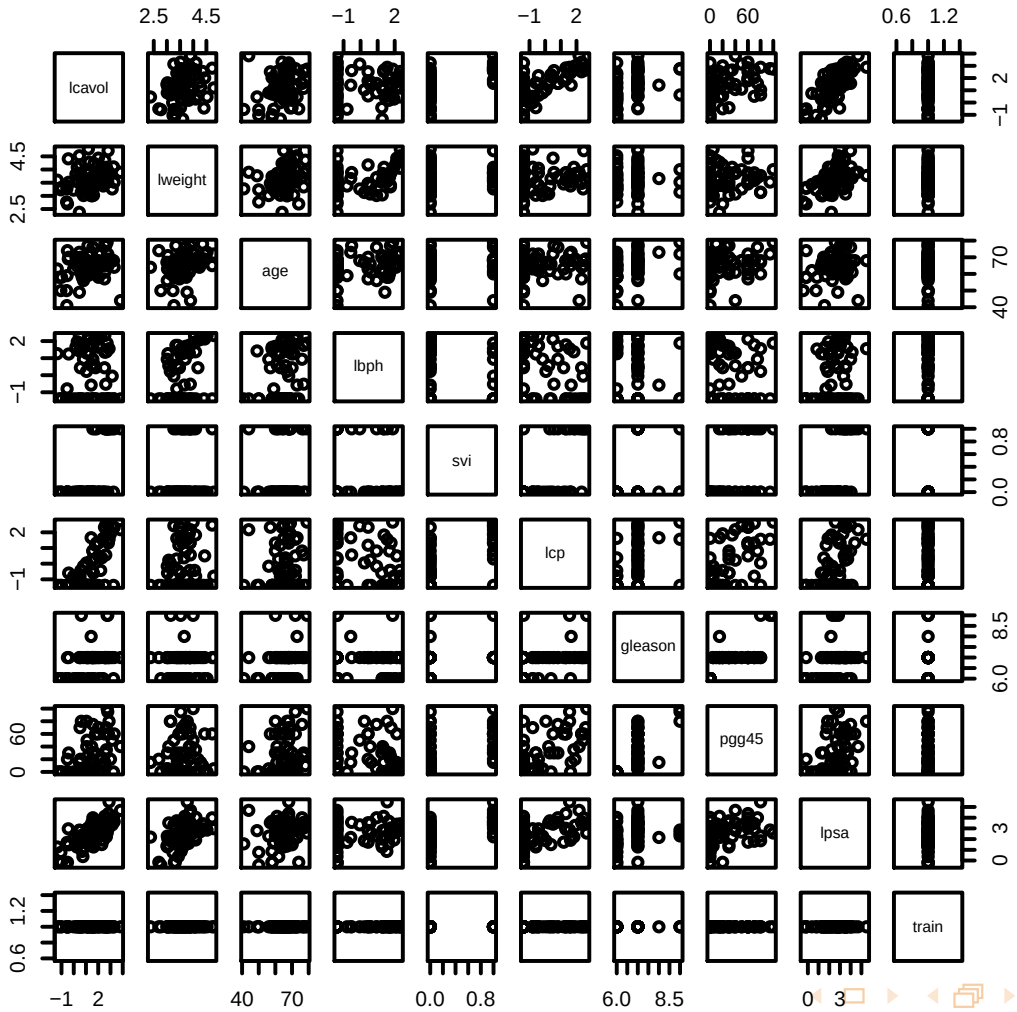
Example We consider data from a study examining the relationship between prostate-specific antigen (PSA) and a number of clinical measures in a sample of men

The explanatory variables are:

- lcavol: Log cancer volume
- lweight: Log prostate weight
- age
- lbph: Log benign prostatic hyperplasia
- svi: Seminal vesicle invasion
- lcp: Log capsular penetration
- gleason: Gleason score
- pgg45: percent of Gleason score 4 or 5

Ridge regression

Example We consider data from a study examining the relationship between prostate-specific antigen (PSA) and a number of clinical measures in a training sample of $n = 67$ men

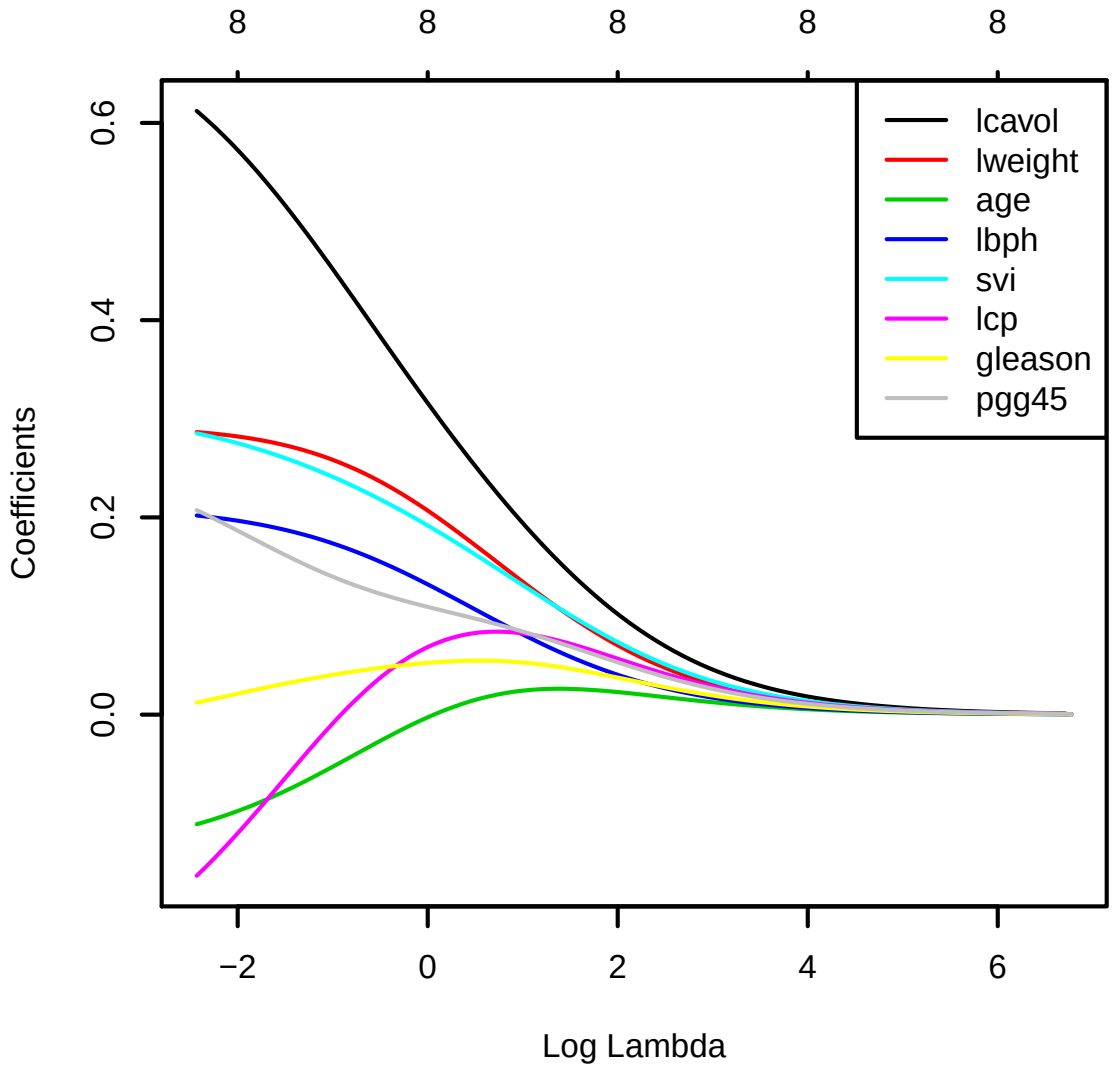


Ridge regression

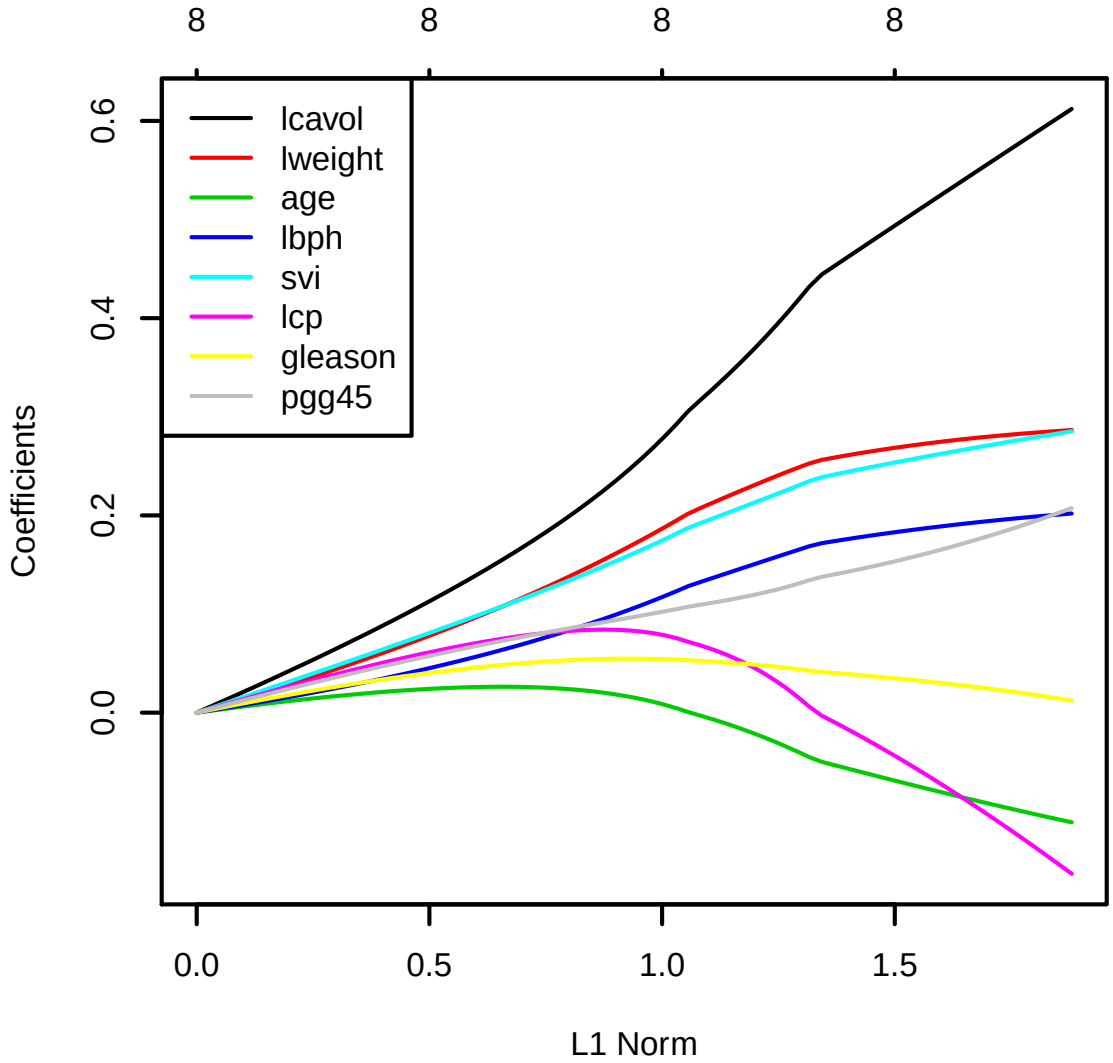
Example We consider data from a study examining the relationship between prostate-specific antigen (PSA) and a number of clinical measures in a training sample of $n = 67$ men

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  2.45234509 0.08701959 28.1815274 1.537669e-35
## xlcavol      0.71640701 0.13350135  5.3662905 1.469415e-06
## xlweight     0.29264240 0.10638488  2.7507894 7.917895e-03
## xage        -0.14254963 0.10211957 -1.3959090 1.680626e-01
## xlbph        0.21200760 0.10312428  2.0558456 4.430784e-02
## xsvi         0.30961953 0.12538985  2.4692552 1.650539e-02
## xlcp        -0.28900562 0.15480404 -1.8669126 6.697085e-02
## xgleason    -0.02091352 0.14257805 -0.1466812 8.838923e-01
## xpgg45       0.27734595 0.15959237  1.7378397 8.754628e-02
```

Ridge regression

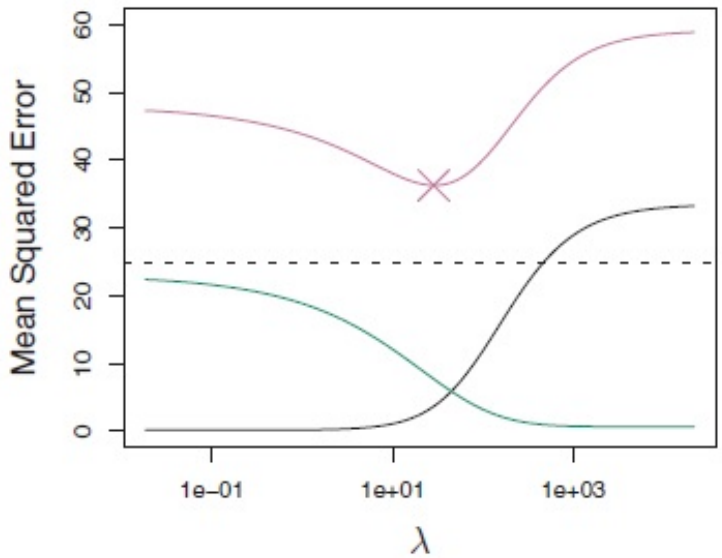


Ridge regression



Ridge Regression vs. Least Squares

- Ridge regression's advantage over least squares is rooted in the bias-variance trade-off
- As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias



Squared bias (black), variance (green) and test mean squared error (purple) for ridge regression predictions on a simulated data set as a function of λ

Image from "An Introduction to Statistical Learning with application in R"

Lasso

- The linear regression model assumes that

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon}$$

  where ϵ is a zero-mean error term
- Given a training sample $(y_1, x_{11}, \ldots, x_{1p}), \ldots, (y_n, x_{n1}, \ldots, x_{np})$, the ordinary least squares estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ solve the problem

$$\boxed{\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2}$$

- The lasso estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ solve the problem

$$\boxed{\begin{array}{ll} \underset{\boldsymbol{\beta}}{\text{minimize}} & \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \\ \text{subject to} & \sum_{j=1}^{p} \mid \beta_j \mid \leq s. \end{array}}$$

Lasso

- The linear regression model assumes that

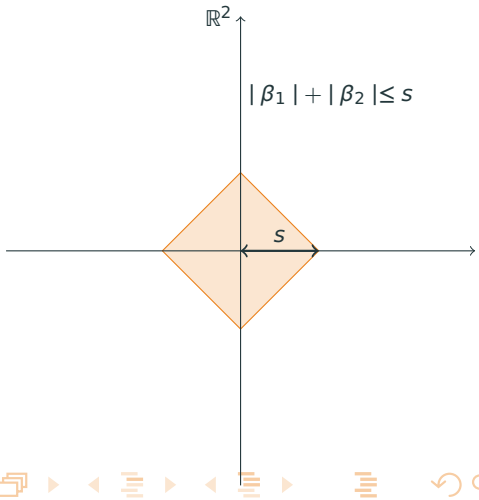
$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon}$$

  where ϵ is a zero-mean error term

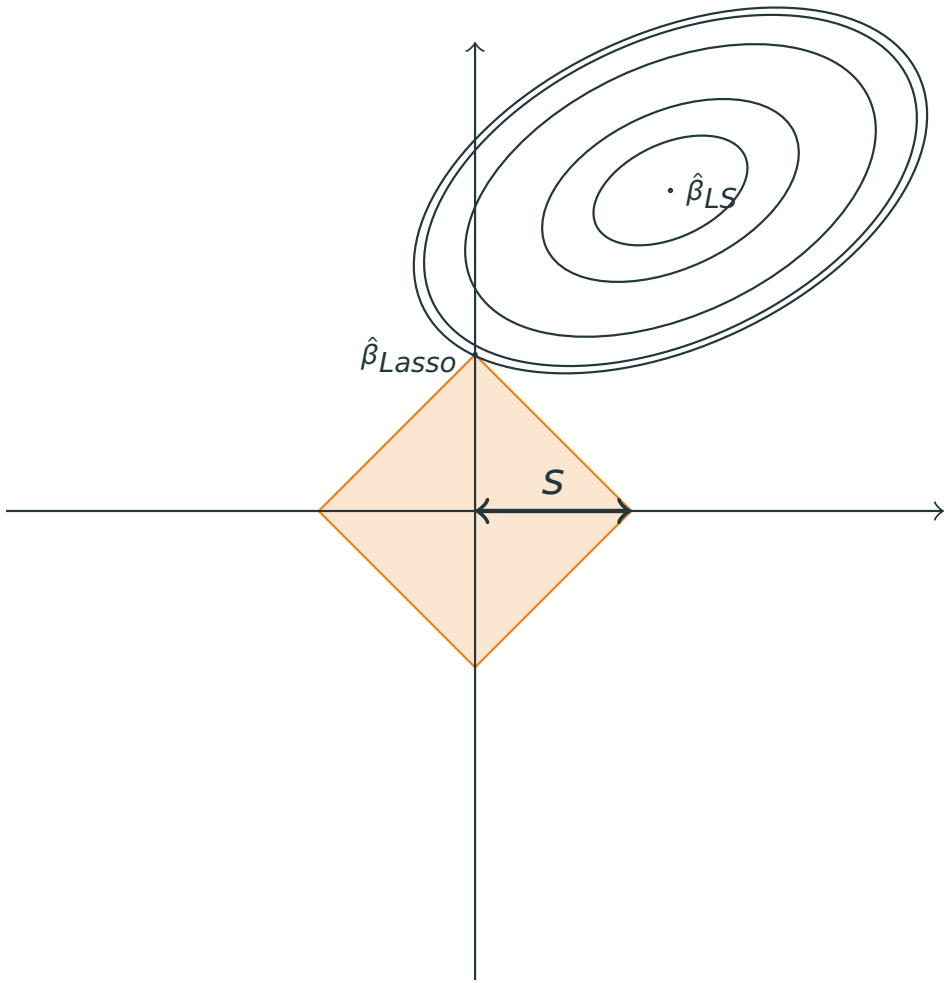
- The lasso estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ solve the problem

$$\boxed{\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} |\beta_j| \leq s. \end{aligned}}$$

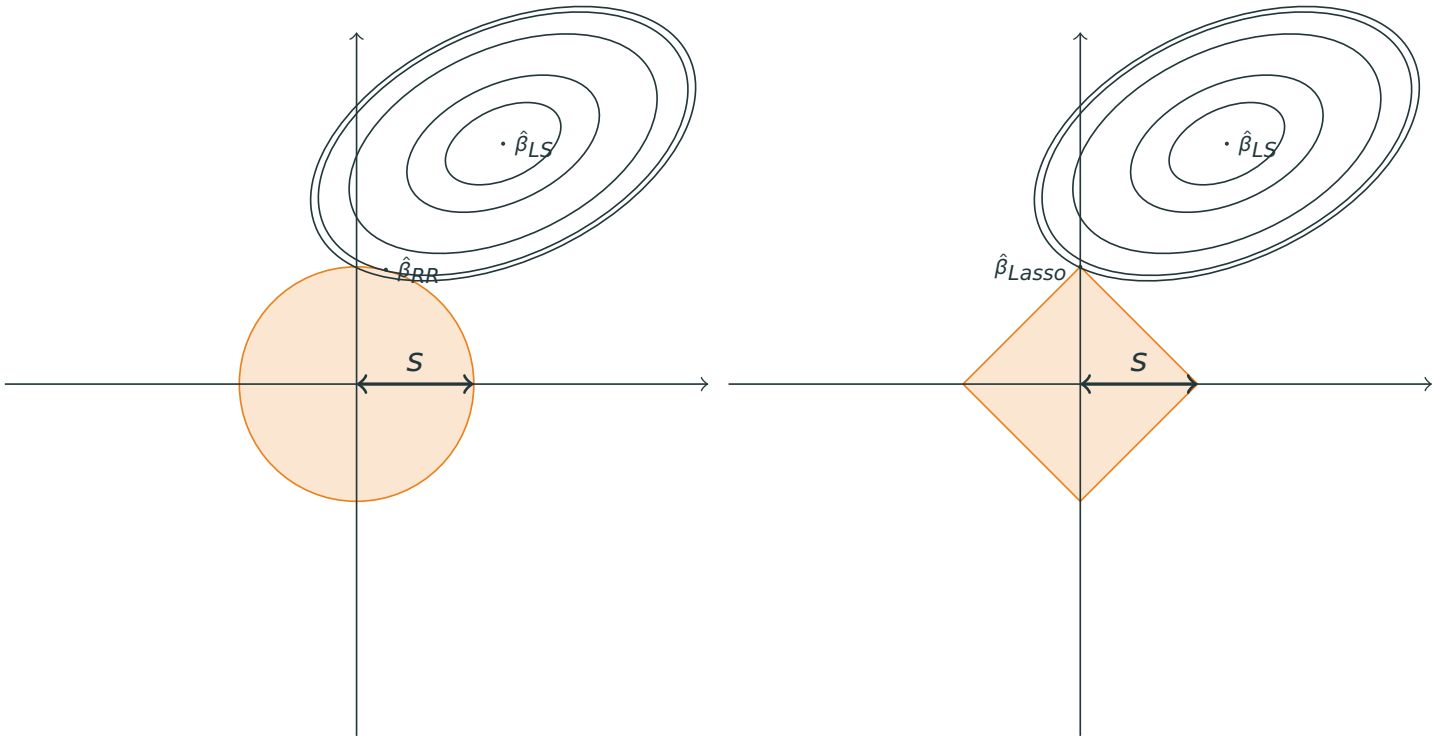
When $p = 2$, the lasso coefficient estimates have the smallest RSS out of all points that lie within the diamond defined by $|\beta_1| + |\beta_2| \leq s$



Lasso



Ridge regression vs. Lasso



Lasso

- The linear regression model assumes that

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon}$$

  where ϵ is a zero-mean error term

- The lasso estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ solve the problem

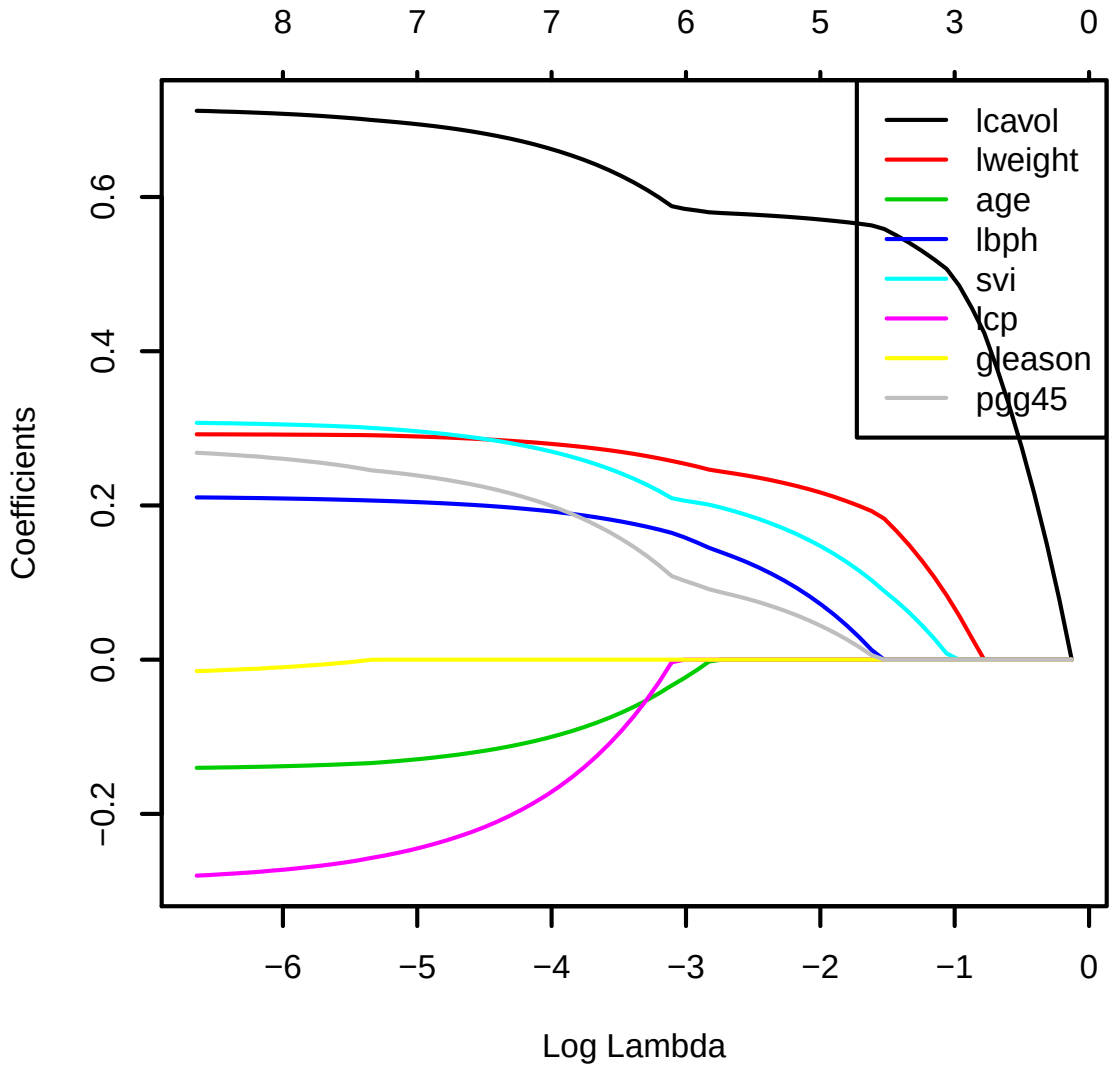
$$\boxed{\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} |\beta_j| \leq s. \end{aligned}}$$

- By the Lagrange multipliers method, the problem can be shown to be equivalent to:

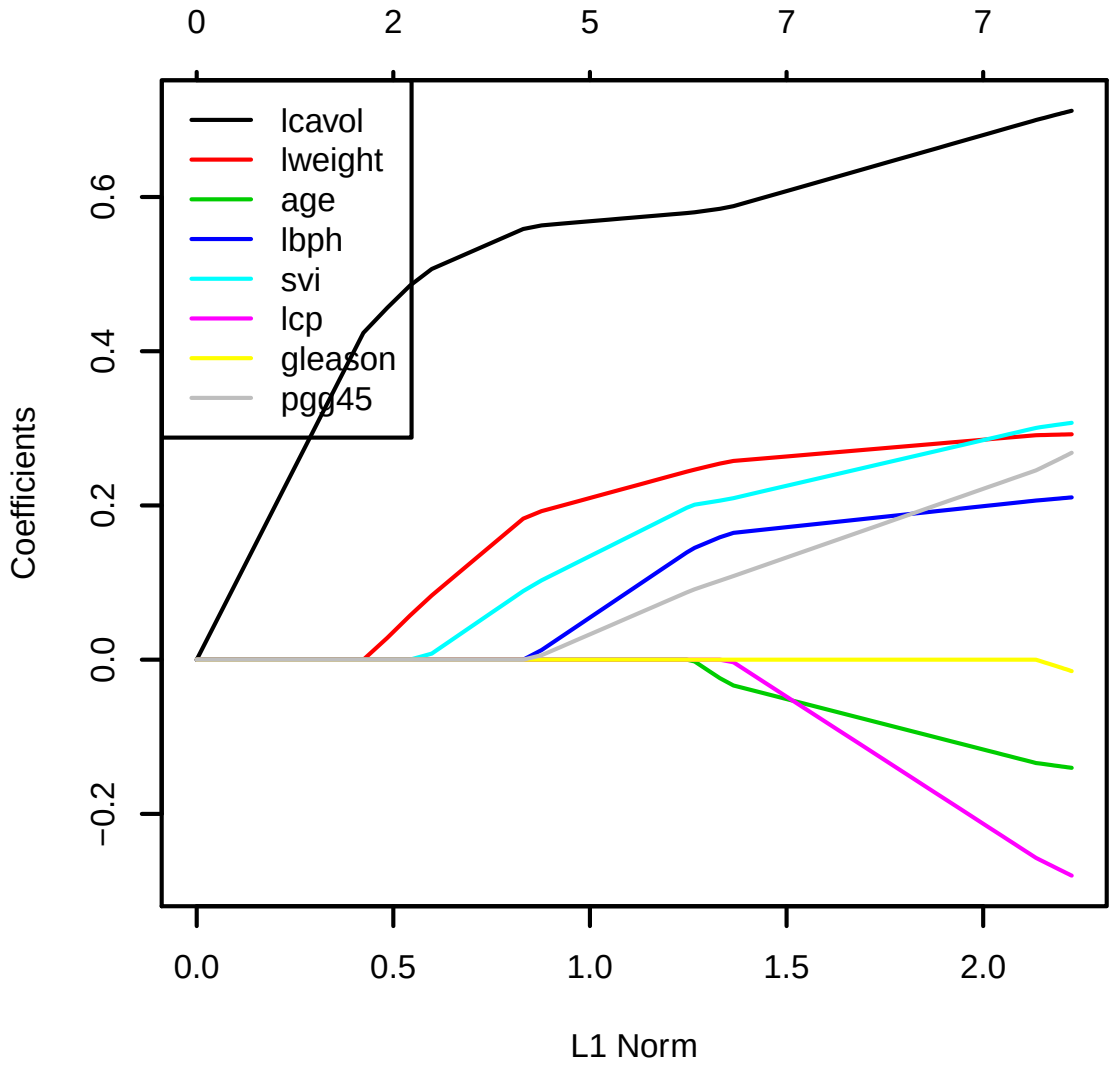
$$\boxed{\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} |\beta_j| .}$$

  (for every value of s there is some λ such that both optimization problems will give the same coefficient estimates)

Lasso

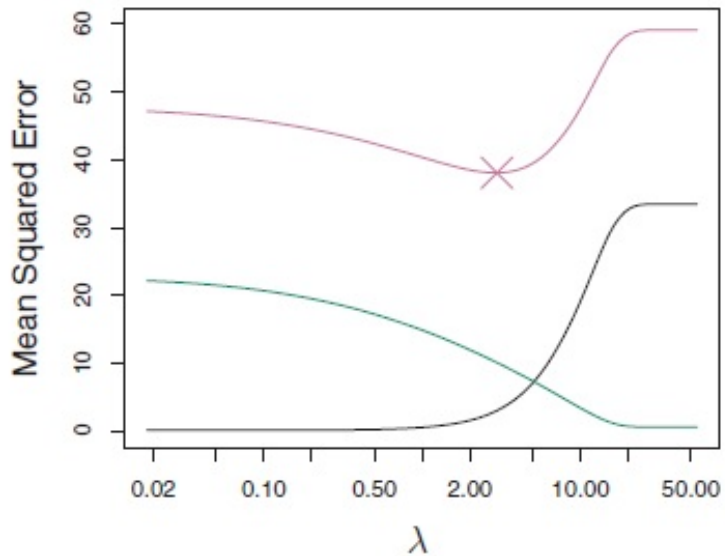

8
7
7
6
5
3
0
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
Coefficients
0.6
0.4
0.2
0.0
−0.2
−6
−5
−4
−3
−2
−1
0
Log Lambda

Lasso


0
2
5
7
7
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
Coefficients
0.6
0.4
0.2
0.0
−0.2
0.0
0.5
1.0
1.5
2.0
L1 Norm

Lasso

- The lasso produces simpler and more interpretable models than ridge regression since lasso models involve only a subset of the predictors
- Ridge regression outperforms the lasso in terms of prediction error when all predictors are related to the response
- Regarding predictor error, lasso behaves similar to ridge regression (as λ increases, the variance decreases and the bias increases)
- But neither ridge regression nor the lasso will universally dominate the other



Squared bias (black), variance (green) and test mean squared error (purple) for ridge regression predictions on a simulated data set as a function of λ

Image from "An Introduction to Statistical Learning with application in R"

Selecting λ

- Implementing ridge regression and the lasso requires a method for selecting a value for the tuning parameter λ
- Cross-validation provides a simple way to tackle this problem: select the tuning parameter value for which the cross-validation error is smallest.
  - Cross-validation methods involve dividing the available set of observations into two parts, a training set and a validation set. The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.

Some comments in high dimensions

- In the high-dimensional setting, the multicollinearity problem is extreme
  - We can never know exactly which variables (if any) truly are predictive of the outcome
  - At most, we can hope to assign large regression coefficients to variables that are correlated with the variables that truly are predictive of the outcome
- It is also important to be particularly careful in reporting errors and measures of model fit in the high-dimensional setting
  - One should never use sum of squared errors, p-values, R^2,... as evidence of a good model fit
  - It is important to instead report results on an independent test set, or cross-validation errors