Regression, Ridge Regression, Lasso

Fabio G. Cozman - fgcozman@usp.br

October 9, 2019

A general definition

- Regression studies the relationship between a *response* variable Y and *covariates* $X_1, \ldots, X_n$.
  - A covariate is also called a *feature* or a *predictor*.
- The term "regression" is usually employed when Y is a continuous variable.
  - That is, variables are *quantitative* rather than *qualitative*.

Basic model and basic questions

- Model:

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

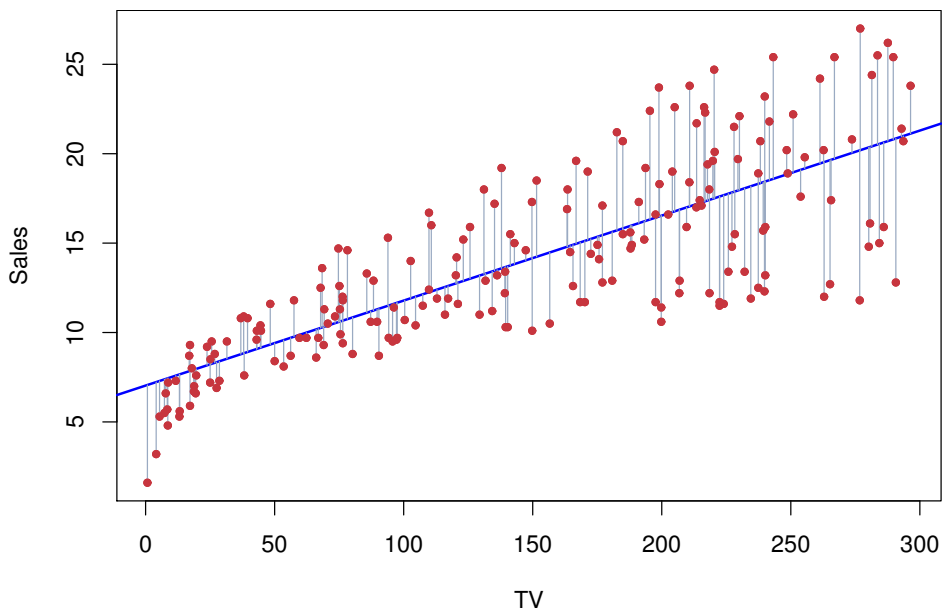
  where $\mathbf{X}$ are the *covariates* and ϵ is some "noise".
- Questions:
  - Is there a relationship? A linear relationship? How strong?
  - Are all covariates useful?
  - How can we predict Y?
  - Usually referred to as *statistical inference*.

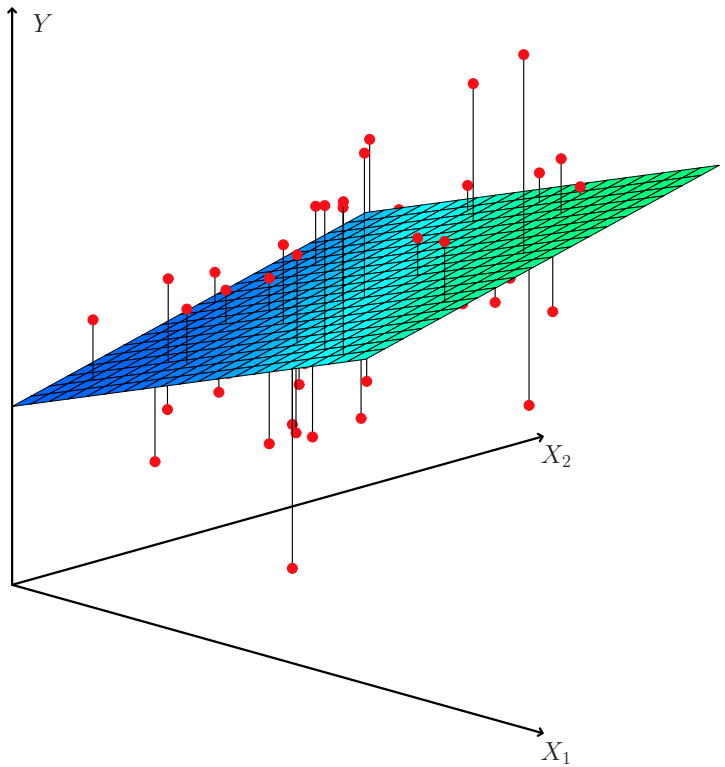
Linear regression

- *Linear regression* adopts a linear relationship:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n.$$

Linear regression: 1D, 2D





The least-squares solution

- Suppose we have observations

$$x_{1,j}, x_{2,j}, \ldots, x_{n,j}, y_j$$

for $j \in \{1, \ldots, N\}$.

The least-squares solution

- Suppose we have observations

$$x_{1,j}, x_{2,j}, \ldots, x_{n,j}, y_j$$

for $j \in \{1, \ldots, N\}$.

- Consider the *residual sum of squares*

$$\mathrm{RSS} = \sum_{j=1}^{n} (y_j - \beta_0 - \beta_1 x_{1,j} - \cdots - \beta_n x_{n,j})^2.$$

The least-squares solution

- Suppose we have observations

$$x_{1,j}, x_{2,j}, \ldots, x_{n,j}, y_j$$

  for $j \in \{1, \ldots, N\}$.
- Consider the *residual sum of squares*

$$\mathrm{RSS} = \sum_{j=1}^{n} (y_j - \beta_0 - \beta_1 x_{1,j} - \cdots - \beta_n x_{n,j})^2.$$

- We might choose $\hat{\beta}_i$ to minimize RSS.

The solution for a single covariate

- Suppose we have

$$Y = \beta_0 + \beta_1 X_1.$$

The solution for a single covariate

- Suppose we have

$$Y = \beta_0 + \beta_1 X_1.$$

- To minimize RSS, we must set:

$$\hat{\beta}_0 = \sum_j y_j/N - \hat{\beta}_1 \sum_j x_{1,j}/N,$$

$$\hat{\beta}_1 = \frac{\sum_j (x_{1,j} - (\sum_k x_{1,k}/N))(y_j - (\sum_k y_k/N))}{\sum_j (x_{1,j} - (\sum_k x_{1,k}/N))^2}.$$

Often used notation

- Adopt $\bar{x}_1 = \frac{\sum_{j=1}^{N} x_{1,j}}{N}$ and $\bar{y} = \frac{\sum_{j=1}^{N} y_j}{N}$.
- Then:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1,$$

$$\hat{\beta}_1 = \frac{\sum_j (x_{1,j} - \bar{x}_1)(y_j - \bar{y})}{\sum_j (x_{1,j} - \bar{x}_1)^2}.$$

Deriving the expressions

- Differentiate $\mathrm{RSS} = \sum_{j=1}^{N}(y_j - \beta_0 - \beta_1 x_{1,j})^2$:

$$\frac{\partial \mathrm{RSS}}{\partial \beta_0} = -2\sum_j (y_j - \beta_0 - \beta_1 x_{1,j}),$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta_1} = -2\sum_j x_{1,j}(y_j - \beta_0 - \beta_1 x_{1,j}).$$

- Solve:

$$\bar{y} - \beta_0 - \beta_1 \bar{x_1} = 0, \qquad \bar{x_1 y} - \beta_0 \bar{x_1} - \beta_1 \bar{x_1^2} = 0.$$

- Get:

$$\beta_0 = \bar{y} - \beta_1 \bar{x_1}, \qquad \beta_1 = \frac{\bar{x_1 y} - \bar{x_1}\bar{y}}{\bar{x_1^2} - \bar{x_1}^2}.$$

A probabilistic version

- Suppose we assume

$$Y = \beta_0 + \beta_1 X_1 + Z,$$

where $Z \sim N(0, \sigma^2)$ is a "probabilistic disturbance".

A probabilistic version

- Suppose we assume

$$Y = \beta_0 + \beta_1 X_1 + Z,$$

where $Z \sim N(0, \sigma^2)$ is a "probabilistic disturbance".

- Then the previous $\hat{\beta}_0$ and $\hat{\beta}_1$ are the maximum likelihood estimates of β_0 and β_1, and moreover

$$\hat{\sigma}^2 = \frac{1}{N} \sum_j (y_j - \hat{y}_j)^2$$

maximizes likelihood for

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1,j}.$$

Maximum likelihood versus unbiasedness

- Funny: the maximum likelihood estimator for σ^2 is not unbiased; often the following unbiased estimator is used:

$$\hat{\sigma}^2 = \frac{1}{N-2} \sum_j (y_j - \hat{y}_j)^2.$$

(This estimator is sometimes called RSE.)

Some properties

- Maximum likelihood estimators of β_0 and β_1 are consistent and asymptotic normal.
- There are closed-form expressions for the variance of these estimators and for confidence intervals (details, not covered in this course, in the Textbook, page 66).

Checking whether there is relationship

- There is a standard (asymptotic) test for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.
- Reject H_0 when

$$\left| \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 \sum_j (x_{1,j} - \bar{x}_1)^2}} \right| > z_{\alpha/2}$$

where $\hat{\sigma}^2$ is (usually) the unbiased estimate of σ^2.

And the p-value

- Test is of the form: $T > c$ for T that depends on data, some c.
- So, p-value: probability that T is larger than observed T, for true H_0 (that is, $\beta_1 = 0$).
- Small p-value: evidence against H_0.

Usual measure of fit

- The R^2 statistic

$$R^2 = 1 - \frac{\mathrm{RSS}}{\sum_{j=1}^{N}(y_j - \bar{y})^2}$$

(the proportion of variability of Y that can be explained by covariates, between 0 and 1; higher is better).

Usual measure of fit

- The R^2 statistic

$$R^2 = 1 - \frac{\mathrm{RSS}}{\sum_{j=1}^{N}(y_j - \bar{y})^2}$$

  (the proportion of variability of Y that can be explained by covariates, between 0 and 1; higher is better).
- For one covariate, R^2 is equal to the square of the correlation between X_1 and Y:

Usual measure of fit

- The R^2 statistic

$$R^2 = 1 - \frac{\mathrm{RSS}}{\sum_{j=1}^{N}(y_j - \bar{y})^2}$$

(the proportion of variability of Y that can be explained by covariates, between 0 and 1; higher is better).

- For one covariate, R^2 is equal to the square of the correlation between X_1 and Y:

$$\left(\frac{\sum_j (x_{1,j} - \bar{x}_1)(y_j - \bar{y})}{\sqrt{\sum_j (x_{1,j} - \bar{x}_1)^2} \sqrt{\sum_j (y_j - \bar{y})^2}} \right)^2 .$$

Many covariates

- Now suppose

$$Y = \beta_1 X_1 + \cdots + \beta_n X_n + Z,$$

where $\mathbb{E}[Z] = 0$ (we can take X_1 equal to one to "simulate" a fixed value at covariates zero).

Many covariates

- Now suppose

$$Y = \beta_1 X_1 + \cdots + \beta_n X_n + Z,$$

where $\mathbb{E}[Z] = 0$ (we can take X_1 equal to one to "simulate" a fixed value at covariates zero).

- The RSS is then

$$(C - AB)^T (C - AB),$$

where

$$A = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{n,1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1,N} & x_{2,N} & \cdots & x_{n,N} \end{bmatrix}, B = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}, C = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}.$$

Regression for many covariates

- For

$$Y = \beta_1 X_1 + \cdots + \beta_n X_n + Z,$$

then

$$\hat{B} = (A^T A)^{-1} A^T C.$$

Some properties

- This estimator is consistent.
- If Z has variance σ^2, then $\hat{B}$ has variance $\sigma^2(A^T A)^{-1}$.
- The estimator $\hat{B}$ is approximately Gaussian, for large N.

Checking whether there is relationship

- There is a standard (asymptotic) test for $H_0 : \beta_1 = \cdots = \beta_n = 0$ versus alternative H_1.

Checking whether there is relationship

- There is a standard (asymptotic) test for $H_0 : \beta_1 = \cdots = \beta_n = 0$ versus alternative H_1.
- Test is "Reject H_0 when F-statistic is large".

Checking whether there is relationship

- There is a standard (asymptotic) test for $H_0 : \beta_1 = \cdots = \beta_n = 0$ versus alternative H_1.
- Test is "Reject H_0 when F-statistic is large".
  - The F-statistic:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/n}{\mathrm{TSS}/(N - n - 1)},$$

where TSS is the total sum of squares:

$$\mathrm{TSS} = \sum_{j=1}^{N} (y_j - \bar{y})^2.$$

Checking whether there is relationship

- There is a standard (asymptotic) test for $H_0 : \beta_1 = \cdots = \beta_n = 0$ versus alternative H_1.
- Test is "Reject H_0 when F-statistic is large".
  - The F-statistic:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/n}{\mathrm{TSS}/(N - n - 1)},$$

    where TSS is the total sum of squares:

$$\mathrm{TSS} = \sum_{j=1}^{N} (y_j - \bar{y})^2.$$

  - The F-statistic has an F-distribution...

Checking subset of parameters

- There are also tests to verify whether subsets of parameters are equal to zero.
- Take $H_0 : \beta_{n-m+1} = \beta_{n-m+2} = \cdots = \beta_n = 0$.

Checking subset of parameters

- There are also tests to verify whether subsets of parameters are equal to zero.
- Take $H_0 : \beta_{n-m+1} = \beta_{n-m+2} = \cdots = \beta_n = 0$.
- Reject H_0 when the following F-statistic is "large":

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/m}{\mathrm{TSS}/(N - n - 1)}$$

where RSS_0 is the sum of residual squares for the model containing only $\beta_1, \ldots, \beta_{n-m}$.

Checking subset of parameters

- There are also tests to verify whether subsets of parameters are equal to zero.
- Take $H_0 : \beta_{n-m+1} = \beta_{n-m+2} = \cdots = \beta_n = 0$.
- Reject H_0 when the following F-statistic is "large":

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/m}{\mathrm{TSS}/(N - n - 1)}$$

where RSS_0 is the sum of residual squares for the model containing only $\beta_1, \ldots, \beta_{n-m}$.

- Possible to check each parameter at a time ($m = 1$).

Usual measure of fit

- The R^2 statistic

$$R^2 = 1 - \frac{\mathrm{RSS}}{\sum_{j=1}^{N}(y_j - \bar{y})^2}$$

(the proportion of variability of Y that can be explained by covariates, between 0 and 1; higher is better).

Usual measure of fit

- The R^2 statistic

$$R^2 = 1 - \frac{\mathrm{RSS}}{\sum_{j=1}^{N}(y_j - \bar{y})^2}$$

(the proportion of variability of Y that can be explained by covariates, between 0 and 1; higher is better).

- But: the more covariates, the higher the R^2 statistic.

Feature selection

- Usually we must discard some covariates.
  - Too many covariates lead to small bias but large variance, while too few covariates lead to small variance but large bias (the bias-variance trade-off).
  - In particular, a covariate should not be a linear function of other covariates...
- We must "score" each set of covariates, to choose the best one. A possible score is empirical error (by cross-validation).

The Akaike Information Criterion

- One popular score is the AIC:

$$L_S - |S|,$$

where

  - S is set of covariates in the scored model, and
  - L_S is the log-likelihood of the model with covariates in S, evaluated at the maximum likelihood estimates.

- That is, "goodness of fit - model complexity"

Another score

- The BIC (Bayesian Information Criterion):

$$L_S - (|S|/2) \log N.$$

- For large N, the posterior probability of the scored model is proportional to e^{BIC}, when all possible models get identical probability.

Yet another score

- Mallow's C_p:

$$2|S|\hat{\sigma}^2 + \sum_{j=1}^{N}(\hat{y}_j^S - y_j)^2,$$

  where $\hat{\sigma}^2$ is the estimate of variance with all covariates, while $\hat{y}_j^S$ are produced with covariates in S.
- This score estimates the "training error" that is expected from a model with covariates in S.

Structure search

- Forward stepwise regression: start with no covariates, add one that leads to largest score, then add one that leds to best score, etc.

- Backward stepwise regression: start with all covariates, drop one that leads to largest score, etc.

Structure search

- Forward stepwise regression: start with no covariates, add one that leads to largest score, then add one that leds to best score, etc.
- Backward stepwise regression: start with all covariates, drop one that leads to largest score, etc.
- There are many other search schemes.

Additional details, not covered in this course, in the Textbook, Section 6.1.

Qualitative features in linear regression

- Use some encoding.
- If values are ordered: turn into numbers.
- If not appropriate to do so: *one-hot* encoding.

Other challenges in linear regression

- Relationship is not linear: detect with tests and residual plots.
- Disturbances are correlated or vary with features.
- Outliers and high leverage points (must be tested for, discarded).
- Collinearity (leads to numerical problems and higher variance).

These issues, not covered in this course, are discussed in the Textbook, Section 3.3.3.

Introducing non-linear features

- Suppose we have X_1 and X_2. We can assume:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2,$$

or any other set of functions of X_1 and X_2.

Introducing non-linear features

- Suppose we have X_1 and X_2. We can assume:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2,$$

or any other set of functions of X_1 and X_2.

  - If functions are polynomials of covariates, we have *polynomial regression*.
  - If functions are all produced by a set of transformations, they are referred to as *basis functions*.

Introducing non-linear features

- Suppose we have X_1 and X_2. We can assume:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2,$$

  or any other set of functions of X_1 and X_2.

  - If functions are polynomials of covariates, we have *polynomial regression*.
  - If functions are all produced by a set of transformations, they are referred to as *basis functions*.

- Other functions lead to General Additive Models (not discussed in this course; see Textbook Section 7.7).

Nonparametric regression

- We might assume that Y depends on *splines* (Y is a piecewise but smooth function of covariates).
- Or we might assume that Y is a function of neighboring points (kNN regression):
  1. To obtain $\hat{y}$ corresponding to given covariates, weigh each point in the neighborhood.
  2. Then run weighted regression with those points only.

Nonparametric regression

- We might assume that Y depends on *splines* (Y is a piecewise but smooth function of covariates).

- Or we might assume that Y is a function of neighboring points (kNN regression):
  1. To obtain $\hat{y}$ corresponding to given covariates, weigh each point in the neighborhood.
  2. Then run weighted regression with those points only.

These topics are not covered in this course; they are discussed in Textbook, Sections 7.5 and 7.6, and Section 3.5.

Shrinkage methods: Regularization

- The idea is to penalize the "size" of parameters, to "shrink" them, to reduce variance (but with increase in bias...).
- Useful to reduce overfitting, particularly when there are too many covariates.
- Two main strategies:
  - ridge regression, and
  - lasso (least absolute shrinkage and selection operator).

Ridge regression

- Suppose we minimize

$$\left(\sum_{j=1}^{N} \left(y_j - \sum_{i=1}^{n} \beta_i x_{i,j} \right)^2 \right) + \lambda \sum_{i=1}^{n} \beta_i^2.$$

- The larger the parameter λ, the smaller the values of $\hat{\beta}_i$.
  - Tuning λ: usually by cross-validation.
- Solution is easy (but biased!):

$$\hat{B} = (A^T A + \lambda I)^{-1} A^T C.$$

Standardization

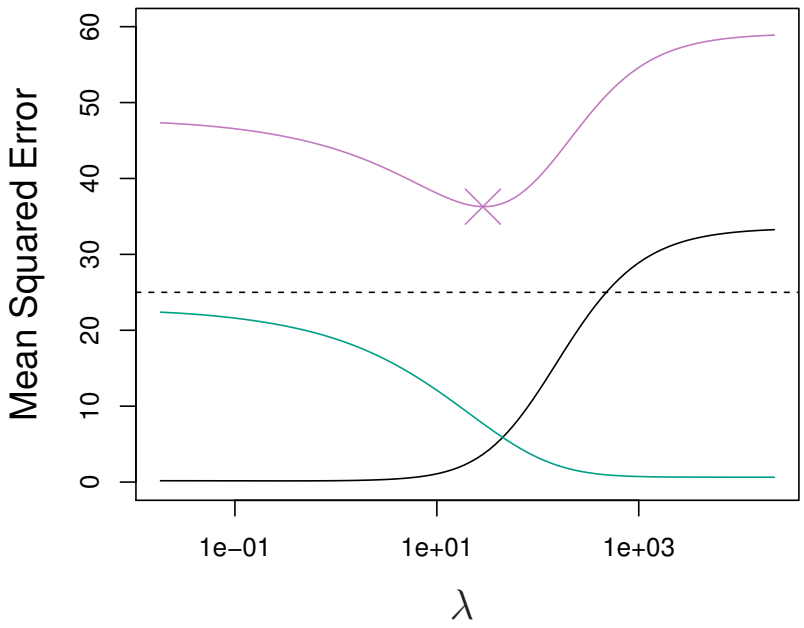
- In ridge regression, the measurement scale for covariates is important...
  - In linear regression, multiplying covariates leads to multiplied estimates.
- Usual assumption: data are standardized (mean 0, variance 1):

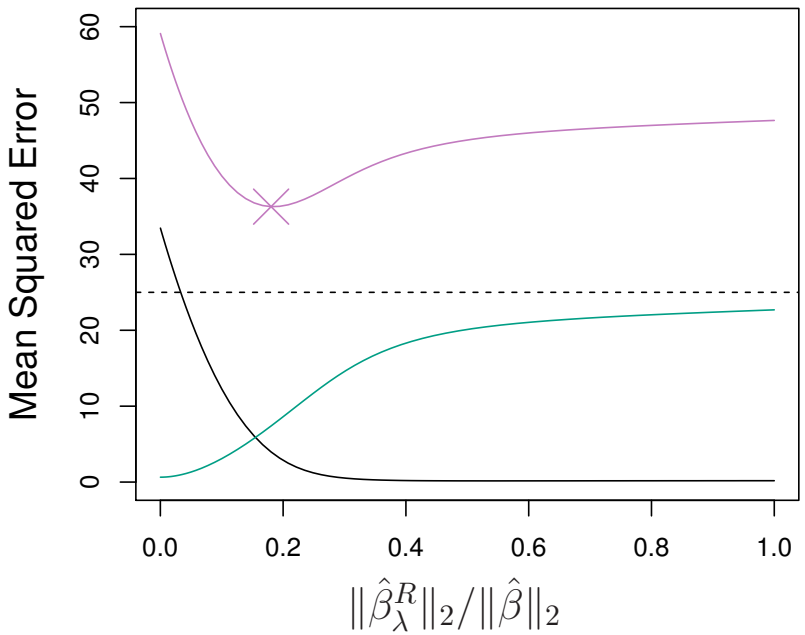
$$\tilde{x}_{i,j} = \frac{x_{i,j}}{\sqrt{(1/N)\sum_{j=1}^{N}(x_{i,j} - \bar{x}_i)^2}}.$$

- Usually Y is centered (mean is subtracted).

The bias-variance trade-off

- Ridge regression increases bias, decreases variance (compared to linear regression).
- So it is useful when variance is large.
  - For instance, when the number of covariates is very large.





The Bayesian interpretation

- Suppose we have Gaussian likelihood and prior proportional to $\mathrm{e}^{-\lambda \sum_i \beta_i^2}$.
- Then the posterior to be maximized (for 0-1 loss) lead to minimization of

$$\left(\sum_{j=1}^{N} \left(y_j - \sum_{i=1}^{n} \beta_i x_{i,j} \right)^2 \right) + \lambda \sum_{i=1}^{n} \beta_i^2.$$

Lasso I

- Suppose we minimize

$$\left(\sum_{j=1}^{N} \left(y_j - \sum_{i=1}^{n} \beta_i x_{i,j} \right)^2 \right) + \lambda \sum_{i=1}^{n} |\beta_i|.$$

- The larger the parameter λ, the smaller the values of $\hat{\beta}_i$.
  - Tuning λ: usually by cross-validation.

Lasso II

- Minimize

$$\left(\sum_{j=1}^{N} \left(y_j - \sum_{i=1}^{n} \beta_i x_{i,j} \right)^2 \right) + \lambda \sum_{i=1}^{n} |\beta_i|.$$

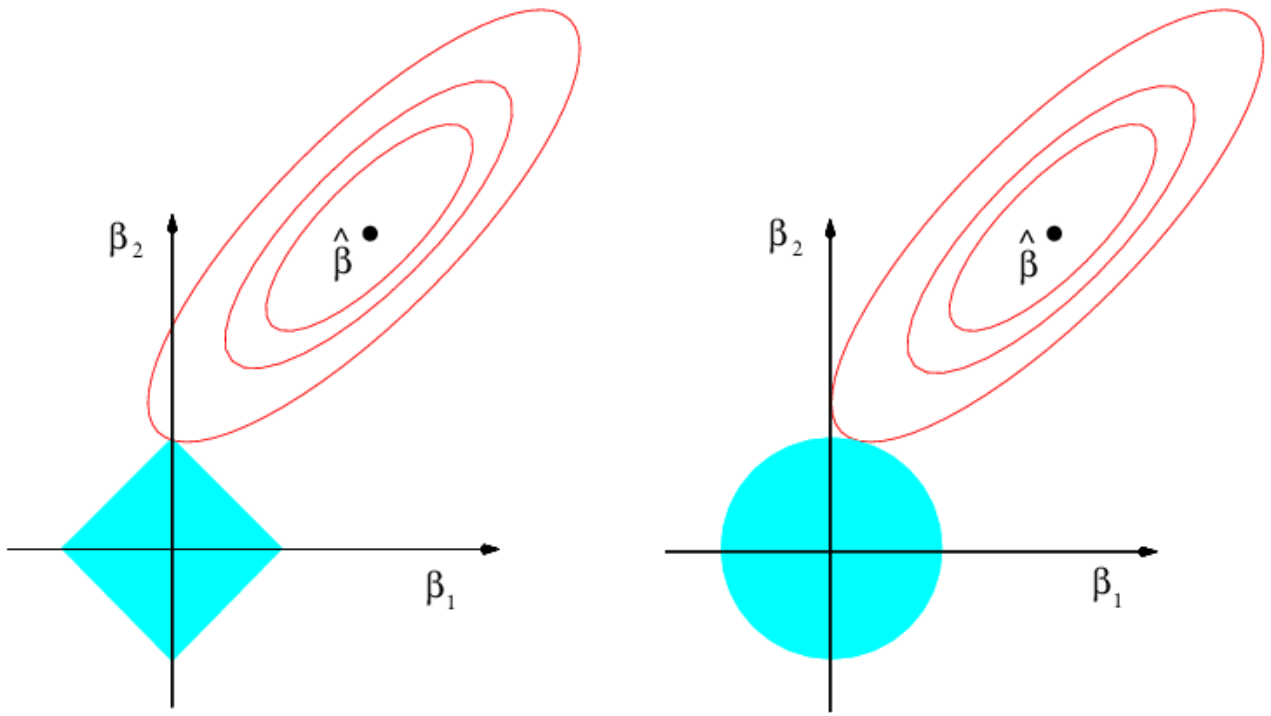
- Usual assumption: X_i standardized (mean 0, variance 1), Y centered.
- Amazing fact: if λ is "large enough", many β_is are set to zero: so covariate selection is done automatically!
  - The result is a sparse solution.

Working with lasso

- No closed-form solution, but convex optimization does it.
- Gains in accuracy have been observed, particularly when too many covariates are present.
  - Too many covariates can overfit any input data...
  - With *too many* covariates, many may have the same predictive power, many may be highly correlated and useless.
  - Thus selection of covariates is a big plus.

Another perspective, with some intuition

- It is possible to show that both ridge regression and lasso minimize RSS, but
  - Ridge regression: subject to $\sum_{i=1}^{n} \beta_i^2 \leq \tilde{\lambda}$.
  - Lasso: subject to $\sum_{i=1}^{n} |\beta_i| \leq \tilde{\lambda}$.



A final note

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.