Linear regression

Xueyi He, Tingyin Wang

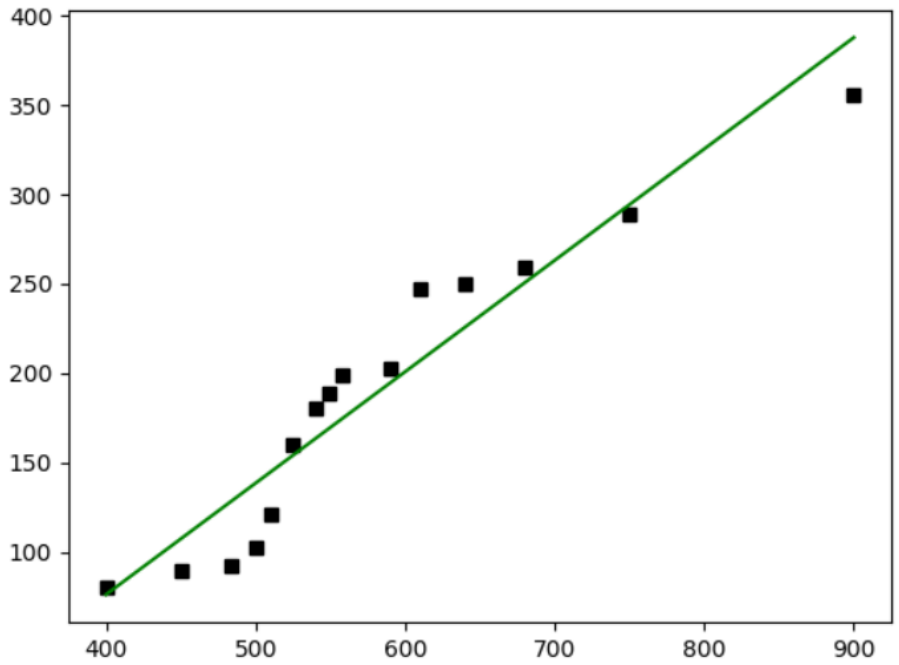
School of Mathematics
Sun Yat-sen University

August 19, 2020

Table of Contents

Table of Contents

Introduction



Given$(x_i, y_i), i = 1, ..., n$, suppose

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $E(\epsilon_i) = 0, var(\epsilon_i) = \sigma^2$. Then,

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x}^2)}, \hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x}.$$

Introduction

Linear regression is the "work horse" of statistics and (supervised) machine learning. When augmented with kernels or other forms of basis function expansion, it can model also non-linear relationships. And when the Gaussian output is replaced with a Bernoulli or multinoulli distribution, it can be used for classification.

Table of Contents

Model specification

- Generative Model:
  This asserts that the response is a linear function of the inputs. This can be written as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}.$$

  where $\boldsymbol{\epsilon}$ is the residual error between our linear predictions and the true response.
  If we assume that ϵ has a normal distribution,

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}),$$

  we have:

$$\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}).$$

  The log likelihood can be written:

$$\ell(\boldsymbol{\theta}) = log\{(2\pi\sigma^2)^{-n/2}\} - \frac{1}{2\sigma^2}(\mathrm{Y} - \boldsymbol{X}\beta)^T(\mathrm{Y} - \boldsymbol{X}\beta).$$

Model specification

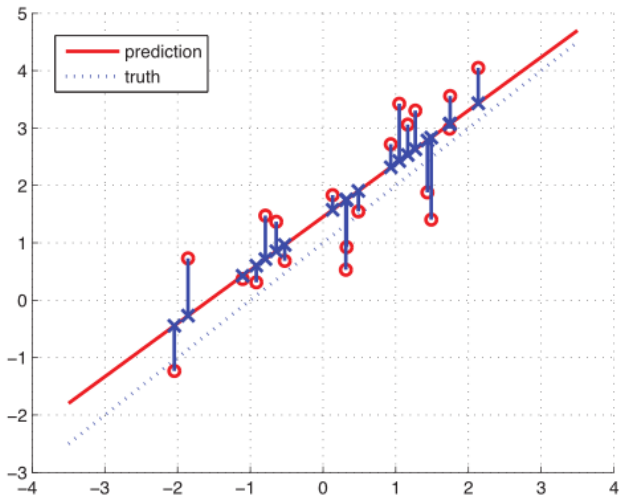
- Generative Model:

$$
\begin{aligned}
\textit{minimize}: Q &= (\mathrm{Y}-\mathrm{X}\boldsymbol{\beta})^T(\mathrm{Y}-\mathrm{X}\boldsymbol{\beta}). \\
Q &= (\mathrm{Y}-\mathrm{X}\boldsymbol{\beta})^T(\mathrm{Y}-\mathrm{X}\boldsymbol{\beta}) \\
&= \mathrm{Y}^T\mathrm{Y}-\mathrm{Y}^T\mathrm{X}\boldsymbol{\beta}-\boldsymbol{\beta}^T\mathrm{X}'\mathrm{Y}+\boldsymbol{\beta}^T\mathrm{X}^T\mathrm{X}\boldsymbol{\beta} \\
&= \mathrm{Y}'\mathrm{Y}-2\mathrm{Y}^T\mathrm{X}\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathrm{X}^T\mathrm{X}\boldsymbol{\beta}, \\
\frac{\partial Q}{\partial \boldsymbol{\beta}} &= -2\mathrm{X}^T\mathrm{Y}+2\mathrm{X}^T\mathrm{X}\boldsymbol{\beta}, \\
\frac{\partial Q}{\partial \boldsymbol{\beta}} &= 0 \Rightarrow \mathrm{X}^T\mathrm{X}\hat{\boldsymbol{\beta}}=\mathrm{X}^T\mathrm{Y}, \\
&\Rightarrow \hat{\boldsymbol{\beta}}=\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathrm{Y}. \qquad (1)
\end{aligned}
$$

Model specification

The corresponding solution $\hat{\beta}$ to this linear system of equations is called the ordinary least squares or OLS solution, which is given by

$$\hat{\beta}_{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}.$$



Model specification

- Discriminative Model:
  We seek a function $f(X)$ for predicting Y given values of the input X.

  This theory requires a loss function $L(Y, f(X))$ for penalizing errors in prediction, and by far the most common and convenient is squared error loss: $L(Y, f(X)) = (Y - f(X))^2$. This leads us to a criterion for choosing f

$$\begin{aligned} \text{EPE}(f) &= \text{E}(Y - f(X))^2 \\ &= \int [y - f(x)]^2 \, \text{Pr}(dx, dy), \end{aligned}$$

  the expected (squared) prediction error .

Model specification

- Discriminative Model:
  By conditioning on X, we can write EPE as

$$\mathrm{EPE}(f) = \mathrm{E}_X \mathrm{E}_{Y|X} \left([Y - f(X)]^2 \mid X \right),$$

  and we see that it suffices to minimize EPE pointwise:

$$f(x) = \mathrm{argmin}_c \, \mathrm{E}_{Y|X} \left([Y - c]^2 \mid X = x \right).$$

  The solution is

$$f(x) = \mathrm{E}(Y \mid X = x),$$

  the conditional expectation, also known as the regression function. Thus the best prediction of Y at any point $X = x$ is the conditional mean, when best is measured by average squared error.

Model specification

- Discriminative Model:
  We settle for
  $$\hat{f}(x) = \text{Ave}\left(y_i \mid x_i \in N_k(x)\right),$$
  where "Ave" denotes average, and $N_k(x)$ is the neighborhood containing the k points in domain closest to x.

  Two approximations are happening here:
  (1) expectation is approximated by averaging over sample data;
  (2) conditioning at a point is relaxed to conditioning on some region "close" to the target point.

  If the linear or some more structured model is appropriate, then we can usually get a more stable estimate than k-nearest neighbors.

Model specification

- Discriminative Model:
  Linear regression assumes that the regression function f(X) is approximately linear in its arguments:

$$f(X) \approx X\beta.$$

  Then, we need to minimize $R(\beta) = E\{(Y - X\beta)^T(Y - X\beta)\}$. The solution is

$$\hat{\beta} = \Big(E\{X^T X\}\Big)^{-1} E\{X^T Y\}. \tag{2}$$

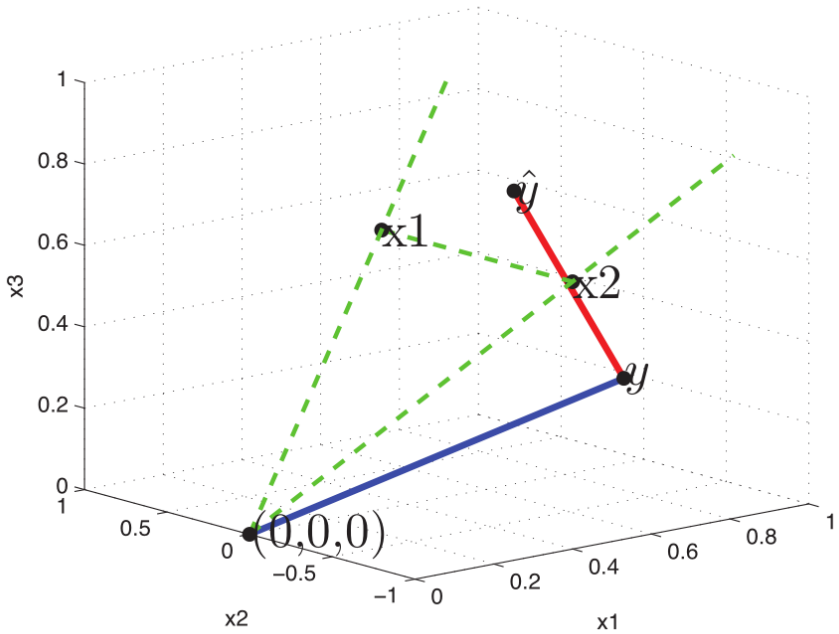
  The least squares solution (1) amounts to replacing the expectation in (2) by averages over the training data.

Geometric interpretation

The corresponding solution $\hat{\beta}$ to this linear system of equations is called the ordinary least squares or OLS solution, which is given by

$$\underset{\hat{Y}\in \text{span}(\{\tilde{x}_1,\ldots,\tilde{x}_p\})}{\text{argmin}} \quad \|Y-\hat{Y}\|_2$$

$$\hat{\boldsymbol{Y}} = X\left(X^T X\right)^{-1} X^T Y$$



Regression by Successive Orthogonalization

Define

$$\langle \mathrm{x}, \mathrm{y} \rangle = \sum_{i=1}^{N} x_i y_i = \mathrm{x}^T \mathrm{y}.$$

Then in the univariate regression analysis,

$$\hat{\boldsymbol{\beta}} = \frac{\langle \mathrm{x}, \mathrm{y} \rangle}{\langle \mathrm{x}, \mathrm{x} \rangle},$$
$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{x}\hat{\boldsymbol{\beta}}.$$

Suppose that the inputs $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_p$ (the columns of the data matrix X)are orthogonal; that is $\langle \mathrm{x}_j, \mathrm{x}_k \rangle = 0$ for all $j \neq k$. Then it is easy to check that the multiple least squares estimates β_j are equal to $\langle \mathrm{x}_j, \mathrm{y} \rangle / \langle \mathrm{x}_j, \mathrm{x}_j \rangle$ (the univariate estimates).

In other words, when the inputs are orthogonal, they have no effect on each other's parameter estimates in the model.

Regression by Successive Orthogonalization

The orthogonalization does not change the subspace spanned by $\{\boldsymbol{x_j}\}, j = 1, ..., p$, it simply produces an orthogonal basis for representing it.

Algorithm 3.1 *Regression by Successive Orthogonalization.*

1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$.
2. For $j = 1, 2, \ldots, p$

   Regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \ldots, , \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{\ell j} = \langle \mathbf{z}_\ell, \mathbf{x}_j \rangle / \langle \mathbf{z}_\ell, \mathbf{z}_\ell \rangle$, $\ell = 0, \ldots, j-1$ and residual vector $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} \mathbf{z}_k$.
3. Regress $\mathbf{y}$ on the residual $\mathbf{z}_p$ to give the estimate $\hat{\beta}_p$.

The result of this algorithm is

$$\hat{\beta}_p = \frac{\langle \mathbf{z}_p, \mathbf{y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}. \tag{3.28}$$

Regression by Successive Orthogonalization

Univariate linear regression: $y_i = a + bx_i + \varepsilon_i$.

Let $\mathrm{x}_0 = [1, 1, \ldots, 1]^T, \mathrm{x}_1$ is the explanatory variable.

Suppose x_0 and x_1 are not orthogonal, then for a simple univariate regression, the solution of a and b:

$$b = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}, a = \frac{1}{n}\sum y_i - b\left(\frac{1}{n}\sum x_i\right).$$

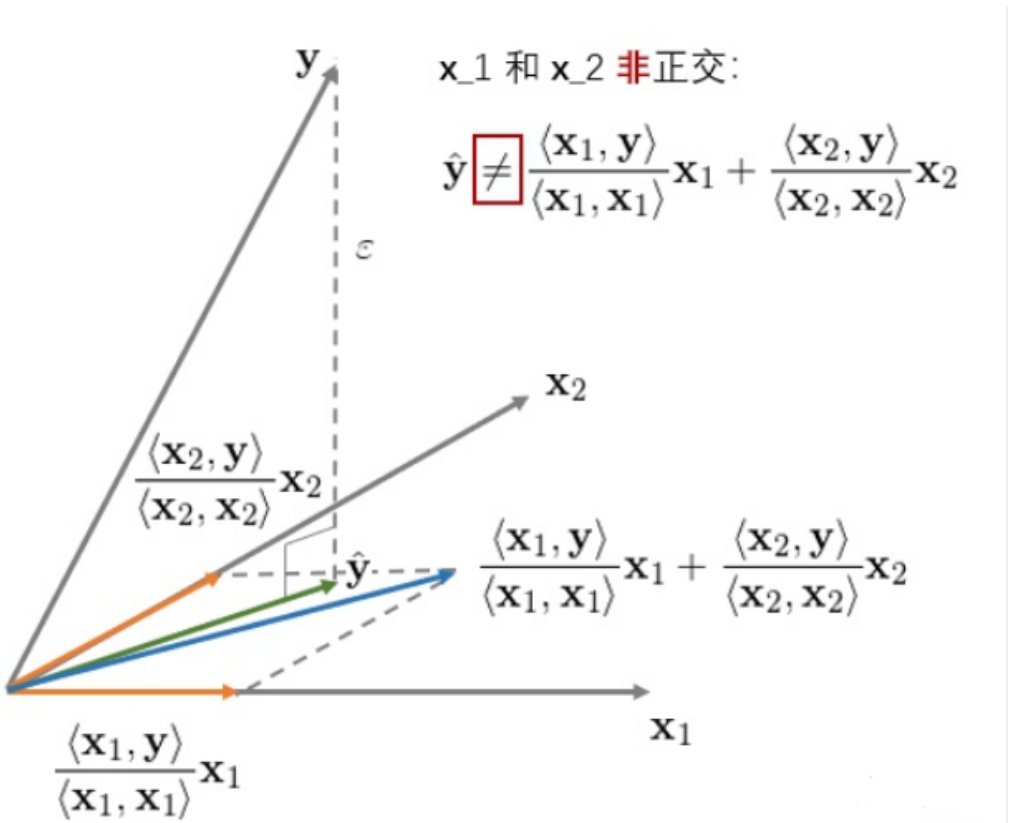
Regression by Successive Orthogonalization

Regression by Successive Orthogonalization: let $z_0 = x_0$, regress x_1 on z_0 to get the residual orthogonal with z_0, denoted as z_1.

$$z_1 = x_1 - \frac{\langle z_0, x_1 \rangle}{\langle z_0, z_0 \rangle} z_0 = x_1 - \bar{x}1,$$

$$\begin{aligned} b &= \frac{\langle z_1, y \rangle}{\langle z_1, z_1 \rangle} = \frac{\langle x_1 - \bar{x}1, y \rangle}{\langle x_1 - \bar{x}1, x_1 - \bar{x}1 \rangle} \\ &= \frac{\sum x_i y_i - \bar{x} \sum y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum \left(x_i^2 - 2 x_i \bar{x} + \bar{x}^2 \right)} \\ &= \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - 2 \left(\sum x_i \right) \bar{x} + n \bar{x}^2} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - 2 n \bar{x}^2 + n \bar{x}^2} \\ &= \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - n \bar{x}^2} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2}. \end{aligned}$$

Regression by Successive Orthogonalization



The multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of x_j on y , after x_j has been adjusted for $x_0, x_1, \ldots, x_{j-1}$ $x_{j+1}, \ldots, x_p$.

Multicollinearity

The estimator of β_p:

$$\hat{\beta}_p = \frac{\langle z_p, y \rangle}{\langle z_p, z_p \rangle}. \tag{3}$$

If x_p is highly correlated with some of the other x_k 's, the residual vector z_p will be close to zero, and from (3) the coefficient $\hat{\beta}_p$ will be very unstable. From (3) we also obtain an alternate formula for the variance estimates $Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \boldsymbol{X})^{-1}\sigma^2$,

$$\text{Var}\left(\hat{\beta}_p\right) = \frac{\sigma^2}{\langle z_p, z_p \rangle} = \frac{\sigma^2}{\|z_p\|^2}.$$

In other words, the precision with which we can estimate $\hat{\beta}_p$ depends on the length of the residual vector z_p; this represents how much of x_p is unexplained by the other x_k' s.

QR decomposition

We can represent step 2 of Algorithm 3.1 in matrix form:

$$\mathbf{X} = \mathbf{Z}\mathbf{\Gamma},$$

where Z has as columns the z_j (in order), and $\mathbf{\Gamma}$ is the upper triangular matrix with entries $\hat{\gamma}_{kj}$. Introducing the diagonal matrix D with j th diagonal entry $D_{jj} = \|z_j\|$, we get

$$\begin{aligned}\mathbf{X} &= \mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\mathbf{\Gamma} \\ &= \mathbf{Q}\mathbf{R},\end{aligned}$$

the so-called QR decomposition of X. Here Q is an $N \times p$ orthogonal matrix, $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, and R is a $p \times p$ upper triangular matrix. The QR decomposition represents a convenient orthogonal basis for the column space of X.

QR decomposition

It is easy to see, for example, that the least squares solution is given by

$$\begin{aligned} \mathrm{QR}\boldsymbol{\beta} &= \mathrm{Y}, \\ \mathrm{R}\boldsymbol{\beta} &= \mathrm{Q}^T\mathrm{Y}, \\ \hat{\boldsymbol{\beta}} &= \mathrm{R}^{-1}\mathrm{Q}^T\mathrm{Y}, \\ \hat{\mathrm{Y}} &= \mathrm{Q}\mathrm{Q}^T\mathrm{Y}. \end{aligned}$$

Equation is easy to solve because R is upper triangular.

Extension

Linear regression can be made to model non-linear relationships by replacing $\boldsymbol{X}$ with some non-linear function of the inputs, $\Phi(\boldsymbol{X})$. That is,

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{I}),$$

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}) = \mathcal{N}(\hat{\boldsymbol{\beta}}\Phi(\boldsymbol{X}), \sigma^2 \boldsymbol{I}).$$

Table of Contents

Some properties

Hat matrix:$\mathrm{H} = \mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T$, symmetric and idempotent.

Distribution of estimator:$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\right)$.

Distribution of fitted value:$\hat{\mathrm{Y}} = \mathrm{HY} \sim N\left(\mathrm{X}\boldsymbol{\beta}, \sigma^2\mathrm{H}\right)$.

Distribution of residuals:$\mathrm{e} \sim N(0, (\mathrm{I}-\mathrm{H})\sigma^2)$.

Analysis of variance:

$$SST = \mathrm{Y}^T\left[\mathrm{I} - \left(\frac{1}{n}\right)\mathrm{J}\right]\mathrm{Y} \sim \sigma^2\chi^2(n-1, \delta_1),$$

$$SSE = \mathrm{Y}^T[\mathrm{I}-\mathrm{H}]\mathrm{Y} \sim \sigma^2\chi^2(n-p, 0),$$

$$SSR = \mathrm{Y}^T\left[\mathrm{H} - \left(\frac{1}{n}\right)\mathrm{J}\right]\mathrm{Y} \sim \sigma^2\chi^2(p-1, \delta_2).$$

Some properties

Hat matrix: $\mathrm{H} = \mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T$.

Property of hat matrix H :

- $\mathrm{HY} = \hat{\mathrm{Y}}, \quad \mathrm{HX} = \mathrm{X}, \quad \mathrm{H}\hat{\mathrm{Y}} = \hat{\mathrm{Y}}, \quad \mathrm{He} = 0$
  $\mathrm{HX} = \mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathrm{X} = \mathrm{X}, \mathrm{H}\hat{\mathrm{Y}} = \mathrm{HX}\hat{\boldsymbol{\beta}} = \mathrm{X}\hat{\boldsymbol{\beta}} = \hat{\mathrm{Y}}.$
- symmetric
  $\mathrm{H}^T = \left(\mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\right)^T = \mathrm{X}(\mathrm{XX})^{-1}\mathrm{X}^T = \mathrm{H}.$
- idempotent
  $\mathrm{HH} = \mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T = \mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T = \mathrm{H}.$

$rank(\boldsymbol{H}) = trace(\boldsymbol{H})$.

Some properties

Distribution of estimator:

$$\hat{\boldsymbol{\beta}} = \left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathrm{Y}.$$

$$\mathrm{Y} \sim N\left(\mathrm{X}\boldsymbol{\beta}, \sigma^2\mathrm{I}\right).$$

$$\begin{aligned} E\{\hat{\boldsymbol{\beta}}\} &= \left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T E\{\mathrm{Y}\} \\ &= \left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathrm{X}\boldsymbol{\beta} = \boldsymbol{\beta}, \end{aligned}$$

$$\begin{aligned} \mathrm{var}\{\hat{\boldsymbol{\beta}}\} &= \left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T \mathrm{var}\{\mathrm{Y}\}\left[\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\right] \\ &= \left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\sigma^2\mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1} = \sigma^2\left(\mathrm{X}^T\mathrm{X}\right)^{-1}, \end{aligned}$$

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\right).$$

Some properties

Distribution of fitted value:

$$\hat{\mathrm{Y}} = \mathrm{X}\hat{\boldsymbol{\beta}} = \mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathrm{Y} = \mathrm{HY}.$$

$$E\{\hat{\mathrm{Y}}\} = E\{\mathrm{HY}\} = \mathrm{H}E\{\mathrm{Y}\} = \mathrm{HX}\boldsymbol{\beta} = \mathrm{X}\boldsymbol{\beta},$$

$$\mathrm{var}\{\hat{\mathrm{Y}}\} = \mathrm{var}\{\mathrm{HY}\} = \mathrm{H}\,\mathrm{var}\{\mathrm{Y}\}\mathrm{H}^T = \sigma^2\mathrm{H},$$

$$\hat{\mathrm{Y}} = \mathrm{HY} \sim N\left(\mathrm{X}\boldsymbol{\beta}, \sigma^2\mathrm{H}\right).$$

Some properties

Distribution of residuals:

$$
\begin{aligned}
\mathrm{e} &= \mathrm{Y} - \hat{\mathrm{Y}} = \mathrm{Y} - \mathrm{HY} = (\mathrm{I} - \mathrm{H})\mathrm{Y}. \\
E\{\mathrm{e}\} &= E\{(\mathrm{I} - \mathrm{H})\mathrm{Y}\} = (\mathrm{I} - \mathrm{H})E\{\mathrm{Y}\} = (\mathrm{I} - \mathrm{H})\mathrm{X}\boldsymbol{\beta} = \mathrm{X}\boldsymbol{\beta} - \mathrm{X}\boldsymbol{\beta} = 0, \\
\mathrm{var}\{\mathrm{e}\} &= (\mathrm{I} - \mathrm{H})\sigma^2\mathrm{I}(\mathrm{I} - \mathrm{H})^T = \sigma^2(\mathrm{I} - \mathrm{H}). \\
\mathrm{e} &\sim N(0, (\mathrm{I} - \mathrm{H})\sigma^2).
\end{aligned}
$$

The matrix $\mathrm{I} - \mathrm{H}$ is also symmetric and idempotent.
$[\mathrm{I} - \mathrm{H}][\mathrm{I} - \mathrm{H}] = \mathrm{I} - \mathrm{H} - \mathrm{H} + \mathrm{H}^2 = \mathrm{I} - \mathrm{H}.$

Some properties

Analysis of variance:

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2.$$

$$SST = \mathrm{Y}^T\mathrm{Y} - \left(\frac{1}{n}\right)\mathrm{Y}^T\mathrm{J}\mathrm{Y} = \mathrm{Y}^T\left[\mathrm{I} - \left(\frac{1}{n}\right)\mathrm{J}\right]\mathrm{Y}, \boldsymbol{J} = 11^T.$$

$$SSE = \mathrm{e}^T\mathrm{e} = [(\mathrm{I}-\mathrm{H})\mathrm{Y}]^T[(\mathrm{I}-\mathrm{H})\mathrm{Y}] = \mathrm{Y}^T(\mathrm{I}-\mathrm{H})^2\mathrm{Y} = \mathrm{Y}^T(\mathrm{I}-\mathrm{H})\mathrm{Y}.$$

$$SSR = SST - SSE$$

$$= \mathrm{Y}^T\left[\mathrm{I} - \left(\frac{1}{n}\right)\mathrm{J}\right]\mathrm{Y} - \mathrm{Y}^T[\mathrm{I}-\mathrm{H}]\mathrm{Y} = \mathrm{Y}^T\left[\mathrm{H} - \left(\frac{1}{n}\right)\mathrm{J}\right]\mathrm{Y}.$$

Some properties

Analysis of variance:

$$SST = \mathrm{Y}^T \left[\mathrm{I} - \left(\frac{1}{n}\right)\mathrm{J}\right]\mathrm{Y}, \qquad \operatorname{rank}\left[\mathrm{I} - \left(\frac{1}{n}\right)\mathrm{J}\right] = n - 1.$$

$$SSE = \mathrm{Y}^T[\mathrm{I} - \mathrm{H}]\mathrm{Y}, \qquad \operatorname{rank}[\mathrm{I} - \mathrm{H}] = n - p.$$

$$SSR = \mathrm{Y}^T \left[\mathrm{H} - \left(\frac{1}{n}\right)\mathrm{J}\right]\mathrm{Y}, \quad \operatorname{rank}\left[\mathrm{H} - \left(\frac{1}{n}\right)\mathrm{J}\right] = p - 1.$$

$\mathrm{H}, \mathrm{J}/\mathrm{n}, \mathrm{I} - \mathrm{J}/\mathrm{n}, \mathrm{I} - \mathrm{H}, \mathrm{H} - \mathrm{J}/\mathrm{n}$ are idempotent and symmetric

$$\operatorname{rank}[\mathrm{H}] = \operatorname{tr}\left[\mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\right] = \operatorname{tr}\left[\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathrm{X}\right] = \operatorname{tr}\left[\mathrm{I}_p\right] = p.$$

$$\operatorname{rank}\left[\mathrm{I} - \left(\frac{1}{n}\right)\mathrm{J}\right] = n-1, \operatorname{rank}[\mathrm{I} - \mathrm{H}] = n-p, \operatorname{rank}\left[\mathrm{H} - \left(\frac{1}{n}\right)\mathrm{J}\right] = p-1.$$

Some properties

Analysis of variance:

$$SST = \mathrm{Y}^T \left[\mathrm{I} - \left(\frac{1}{n}\right)\mathrm{J}\right]\mathrm{Y} \sim \sigma^2\chi^2(n-1,\delta_1),$$

$$SSE = \mathrm{Y}^T[\mathrm{I}-\mathrm{H}]\mathrm{Y} \sim \sigma^2\chi^2(n-p,0),$$

$$SSR = \mathrm{Y}^T \left[\mathrm{H} - \left(\frac{1}{n}\right)\mathrm{J}\right]\mathrm{Y} \sim \sigma^2\chi^2(p-1,\delta_2).$$

Where

$$\delta_1 = \frac{1}{\sigma^2}(\mathrm{X}\boldsymbol{\beta})^T \left[\mathrm{I} - \left(\frac{1}{n}\right)\mathrm{J}\right]\mathrm{X}\boldsymbol{\beta},$$

$$\delta_2 = \frac{1}{\sigma^2}(\mathrm{X}\boldsymbol{\beta})^T \left[\mathrm{H} - \left(\frac{1}{n}\right)\mathrm{J}\right]\mathrm{X}\boldsymbol{\beta}.$$

$SSR \perp SSE$(See Cochran's Theorem in Appendix).

Test whether $\beta_k = 0$

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right).$$

$$\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2),$$

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\sigma} \sim N(0,1),$$

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-p),$$

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-p}},$$

$$t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t(n-p).$$

c_{jj} is jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$.

Test whether $\beta_k = 0$ or several $\beta_k = 0$

$H_0 : \beta_q = \ldots = \beta_{p-1} = 0, H_A$: At least one of $\beta_q \ldots \beta_{p-1} \neq 0$

Full Model:
$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma^2\right)$
Reduced Model:
$Y_i = \beta_0 + \beta_1 X_{i1} \ldots + \beta_{q-1} X_{i,q-1} + \varepsilon_i \quad (q < p).$

Full Model: $\mathrm{SSE}(F) = \mathrm{SSE}\left(X_1, X_2 \ldots, X_{p-1}\right), \quad df_F = n - p.$
Reduced Model: $\mathrm{SSE}(R) = \mathrm{SSE}\left(X_1, X_2 \ldots, X_{q-1}\right), \quad df_R = n - q.$

General Linear Test:

$$F^* = \frac{\left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right]}{\left[\frac{SSE(F)}{df_F}\right]} \overset{H_0}{\sim} F(p - q, n - p).$$

R square and adjusted R square

The coefficient of multiple determination R^2 is defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} (0 \leq R^2 \leq 1).$$

R^2 always increases when there are more variables.

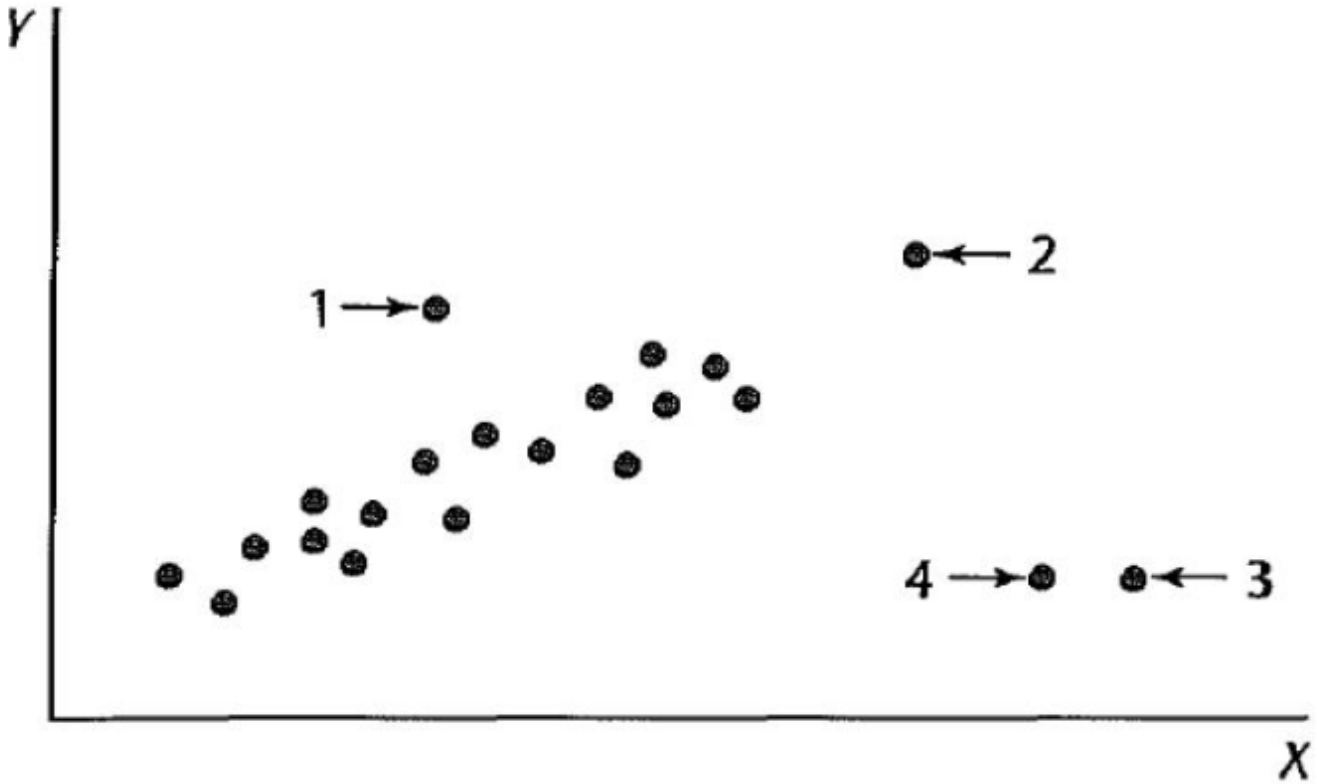
Adjusted R^2 :

$$R_a^2 = 1 - \frac{[SSE/(n-p)]}{[SST/(n-1)]} = 1 - \frac{MSE}{MST} = 1 - \left(\frac{n-1}{n-p}\right) \frac{SSE}{SST}.$$

Adjusted R^2 may decrease when p is large.

Table of Contents

Outliers



Outlying Y Observations

Residuals:

$$\mathrm{e} \sim N(0, (\mathrm{I} - \mathrm{H})\sigma^2)$$

Studentized residual:

$$\frac{e_i}{s\{e_i\}} = \frac{e_i}{\sqrt{MSE\,(1 - h_{ii})}}.$$

$$E\{e_i\} = 0; \quad \text{Let } h_{ij} = (i,j)^{th} \text{ element of } \mathrm{H} = \mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T,$$

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii}), \quad \sigma\{e_i, e_j\} = -h_{ij}\sigma^2, \quad \forall i \neq j$$

$$s^2\{e_i\} = MSE\,(1 - h_{ii}), \quad s\{e_i, e_j\} = -h_{ij}MSE, \quad \forall i \neq j$$

The rough criterion is that the studentized residual is greater than 2 or less than -2.

Outlying X-Cases

$$\mathrm{H} = \mathrm{X}\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T, \quad h_{ij} = \mathrm{x}_i^T\left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{x}_j,$$
$$0 \le h_{ii} \le 1, \quad \sum_{i=1}^{n} h_{ii} = \mathrm{trace}(\mathrm{H}) = p.$$

- h_{ii} known as the leverage of ith case. It is a measure of distance between the X_i value and the mean of the X values.
- Cases with X-levels close to the "center" of the sampled X-levels will have small leverages. Cases with "extreme"levels have large leverages
- Large leverage values: $h_{ii} > 2p/n$.

$$\hat{\mathrm{Y}} = \mathrm{HY} \Rightarrow \hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{i1} Y_1 + h_{i2} Y_2 + \ldots + h_{in} Y_n.$$

- Thus h_{ii} is a measure of how much Y_i is contributing to the prediction $\hat{Y}_i$.
- Cases with large leverages have the potential to "pull" the regression equation toward their observed Y-values.

Identifying Influential Cases

Influence on all fitted values (Cook's Distance):

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{pMSE} = \frac{\left(\hat{\mathrm{Y}} - \hat{\mathrm{Y}}_{(i)}\right)^T\left(\hat{\mathrm{Y}} - \hat{\mathrm{Y}}_{(i)}\right)}{pMSE}$$
$$= \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] = \frac{h_{ii}}{(1-h_{ii})}\frac{\tilde{e}_i^2}{p}.\text{(See Appendix)}$$

where $\hat{Y}_{j(i)}$ is the fitted value when regression is fit on the other n-1 cases, $\tilde{e}_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$ is studentized residual.
Problem cases are $> F(0.50; p, n-p)$.

Robust linear regression

One way to achieve robustness to outliers is to replace the Gaussian distribution for the response variable with a distribution that has heavy tails. Such a distribution will assign higher likelihood to outliers, without having to perturb the straight line to "explain" them.

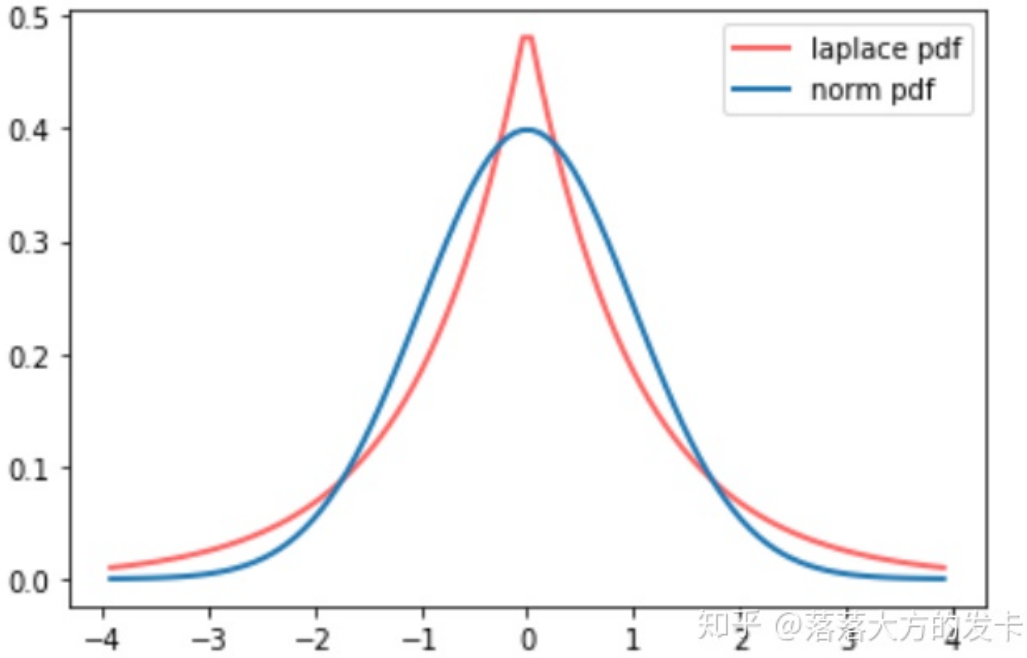
Gaussian distribution:

$$p(Y|\boldsymbol{X}, \beta) = \mathcal{N}(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}).$$

Laplace distribution:

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{X}, \beta, b) &= Lap(\boldsymbol{Y}|\beta^T \boldsymbol{X}, b) \\ &\propto exp(-\frac{1}{b}|\boldsymbol{Y} - \boldsymbol{X}\beta|). \end{aligned}$$

Robust linear regression



The robustness arises from the use of $|\boldsymbol{Y} - \boldsymbol{X}\beta|$ instead of $(\boldsymbol{Y} - \boldsymbol{X}\beta)'(\boldsymbol{Y} - \boldsymbol{X}\beta)$. For simplicity, we will assume b is fixed. Let $r_i \triangleq y_i - \beta^T \boldsymbol{x_i}$ be the $i'th$ residual. The NLL(negative log likelihood) has the form

$$\ell(\beta) = \sum_i |r_i(\beta)|.$$

Robust linear regression

This is a non-linear objective function.
But we can convert the NLL to a linear objective, subject to linear constraints, using the following split variable trick.
First we define

$$r_i^+ = \frac{1}{2}(|r_i| + r_i) \geq 0, r_i^- = \frac{1}{2}(|r_i| - r_i) \geq 0, |r_i| \triangleq r_i^+ + r_i^-,$$

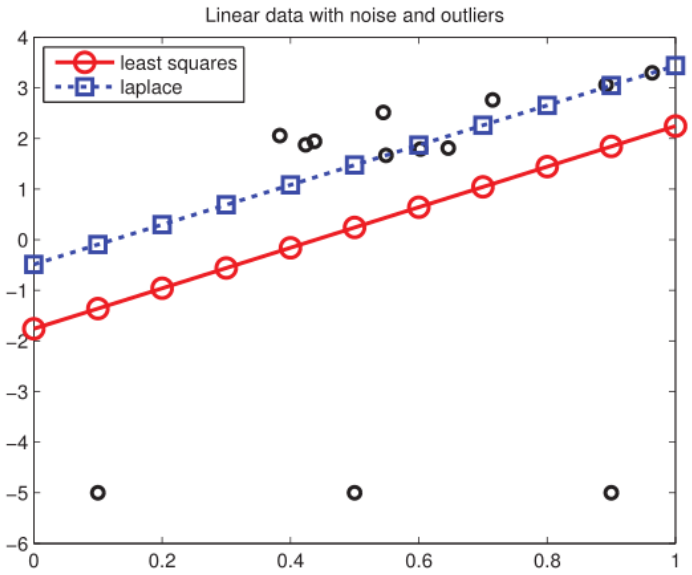
and then we impose the linear inequality constraints that $r_i^+ \geq 0$, $r_i^- \geq 0$.
Now the constrained objective becomes

$$\min_{\boldsymbol{w}, \boldsymbol{r}^+, \boldsymbol{r}^-} \sum_i (r_i^+ + r_i^-) s.t. r_i^+ \geq 0, r_i^- \geq 0, \boldsymbol{\beta}^T \boldsymbol{x_i} + r_i^+ - r_i^- = y_i.$$

This is an example of a linear program with p + 2N unknowns and 3N constraints. It is a convex optimization problem, it has a unique solution solved by any Linear Programming solver.

Outlying Y Observations

If we have outliers in our data, this can result in a poor fit.



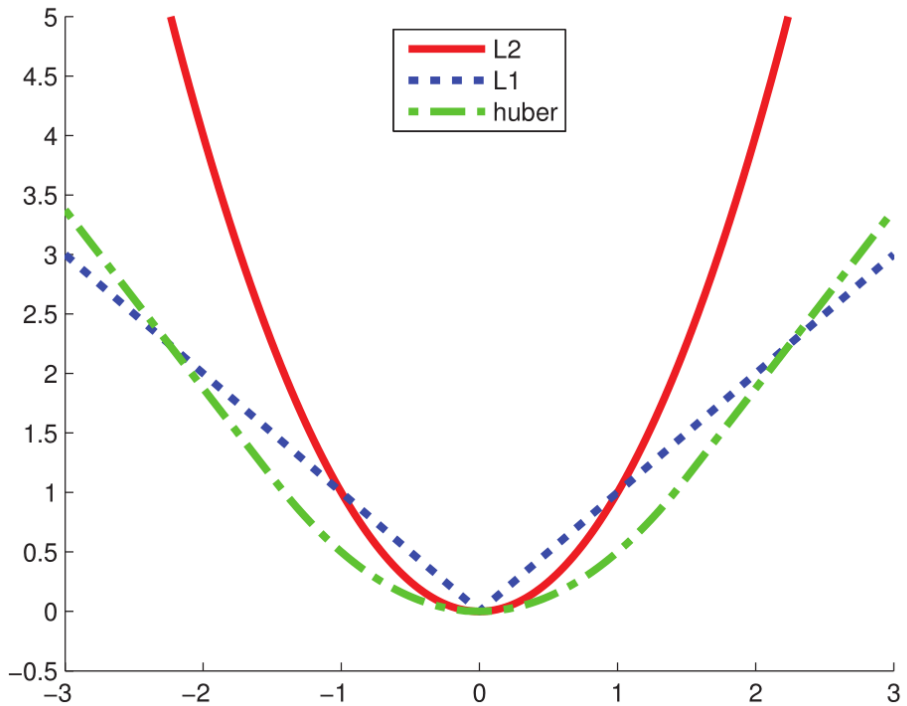
(a)

This is because squared error penalizes deviations quadratically, so points far from the line have more affect on the fit than points near to the line.

Robust linear regression

An alternative to using NLL under a Laplace likelihood is to minimize the Huber loss function (Huber 1964), defined as follows:

$$L_H(r, \delta) = \begin{cases} r^2/2 & \text{if } |r| \leq \delta \\ \delta|r| - \delta^2/2 & \text{if } |r| > \delta \end{cases}$$



R example

```
> library(MASS)
> names(Boston)
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"     "rad"
[10] "tax"     "ptratio" "black"   "lstat"   "medv"
> lm.fit = lm(medv ~ ., data = Boston)
> summary(lm.fit)

Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

R example

```r
par(mfrow=c(2,2))
plot(lm.fit)
```

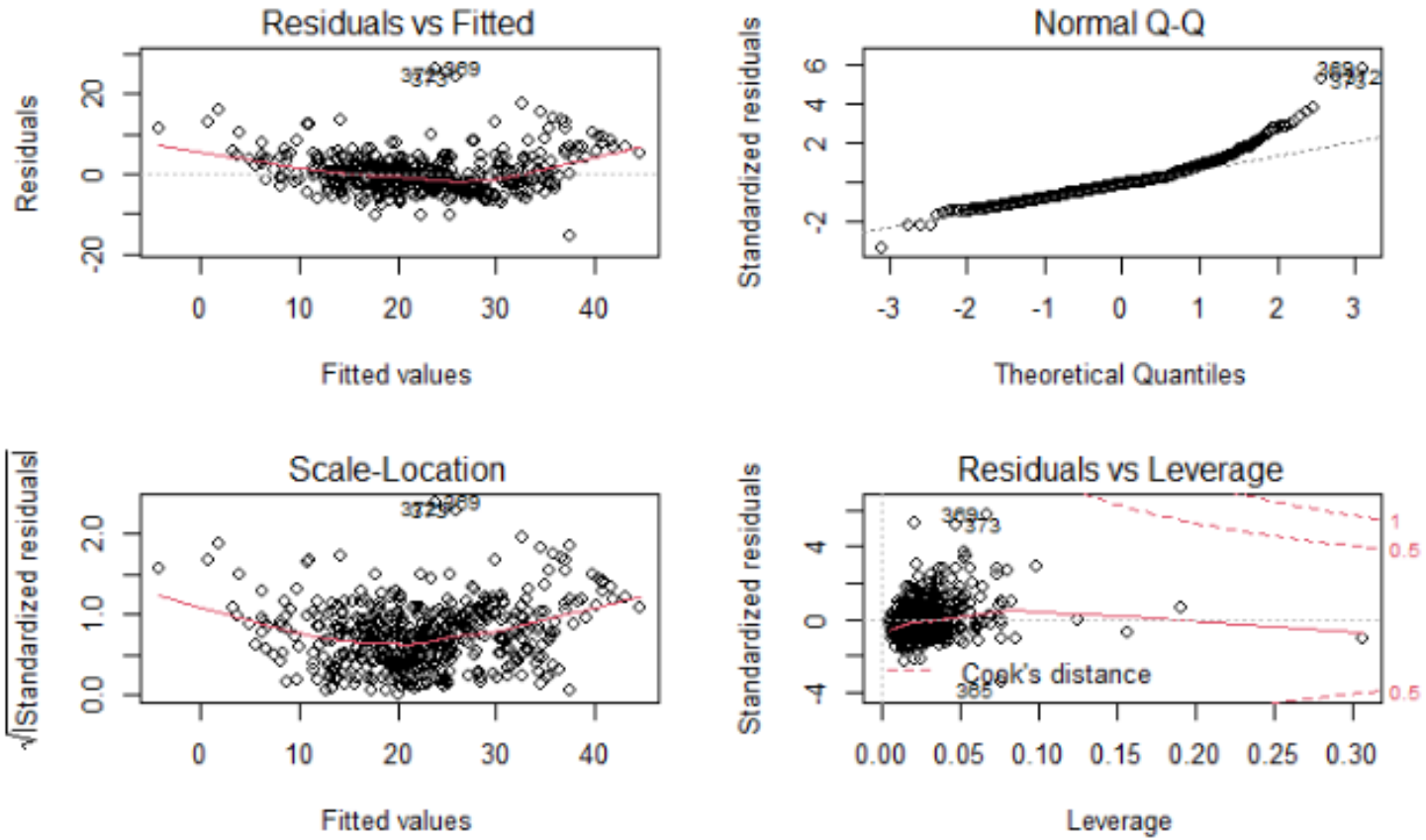


Table of Contents

Too Much Variables

Some variables does not contain enough useful information and can be redundant.An example: A hospital is trying to study the survival time after a liver resection. 10 variables of 54 patients are collected, but we only take 6 of them into consideration.

Variables

V1	Blood clotting （凝血）score
V2	Prognostic index（预后指数）
V3	Enzyme function test score （酶功能测试分数）
V4	Liver function test score （肝功能测试分数）
V5	Age （年龄）
V10 (considered as Y)	Survival time（存活时间） (log-form)

Models

Empty Model	$V_{10} \sim null$
Model 1	$V_{10} \sim V_1 + V_2 + V_3$
Model 2	$V_{10} \sim V_1 + V_2 + V_3 + V4 + V5$

Error

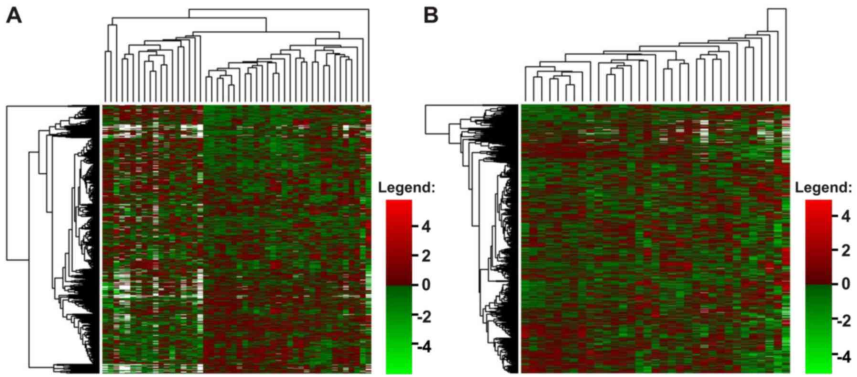
NO. of Variables	0	3	5
Predicted Error	0.147569	0.075036	0.071955

High Dimensional Problem

Sometimes extreme cases occur and the amount of variables p can be so large and surpass sample size n. And a problem like this is unsolvable (basically because $X^T X$ is not invertible).

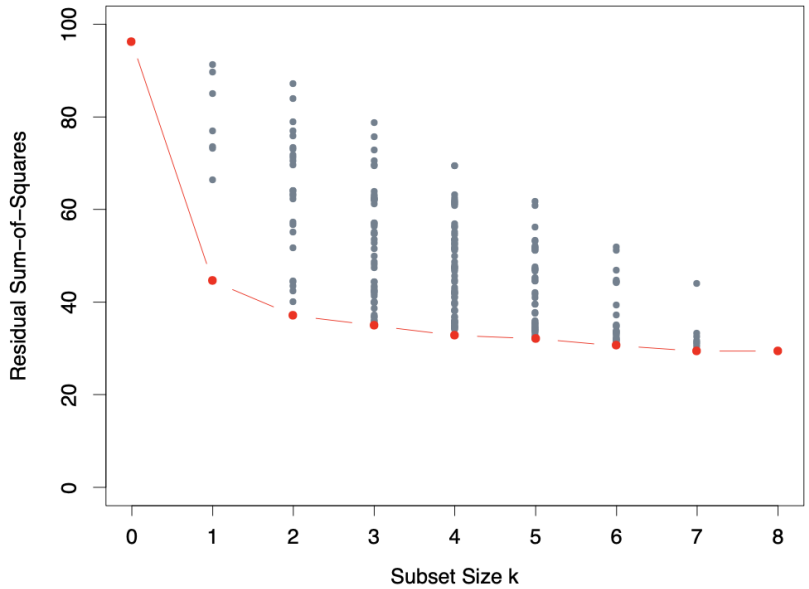
Example: Gene expression pattern of a certain kind of disease

- The amount of genes are simply too large but the NO. of patients included in research is limited.
- For practical use, researchers want to find a few genes that are highly correlated with the disease.



Intuitive Thought

Intuitively, to do subset selection, assuming there are p variables in total, we can do regression with k variables for each $k \in \{0, 1, 2, \cdots, p\}$(when $n > p$),($k \in \{0, 1, 2, \cdots, p\}$(when $n < p$) find the best one for each k . However, the problem arises when we try to define the best 'k'.



Note that we cannot use the RSS to determine the right k because it is always decreasing when k increase.

Model Selection Methods

We change our choice of λ to create different criteria: Some criteria that are commonly used

$$MSE_p + \lambda p \frac{\hat{\sigma}^2}{n},$$

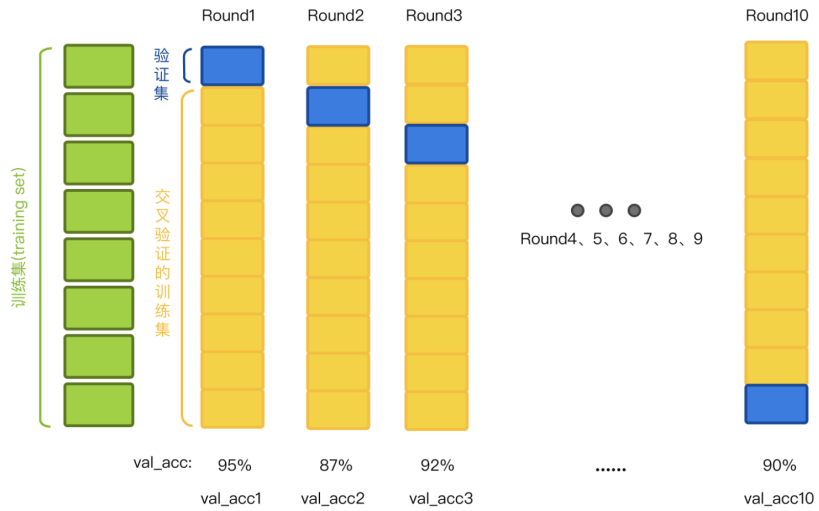
- Akaike Information Criterion(AIC), given as $AIC = 2p - 2\ln(\hat{L})$, under the assumption that the noise satisfies Gaussian distribution, is a coefficient 'away' of taking $\lambda = 2$ in the above equation;
- Bayesian Information Criterion(BIC),also given as $BIC = p\ln(n) - 2\ln(\hat{L})$, under the assumption that the noise satisfies Gaussian distribution, is a coefficient 'away' of taking $\lambda = \ln(n)$ in the above equation;

Model Selection Methods

Cross Validation

Cross-validation is a model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.

The most commonly used CV is called k-fold cross-validation,



Subset Selection Methods

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots p$:

   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

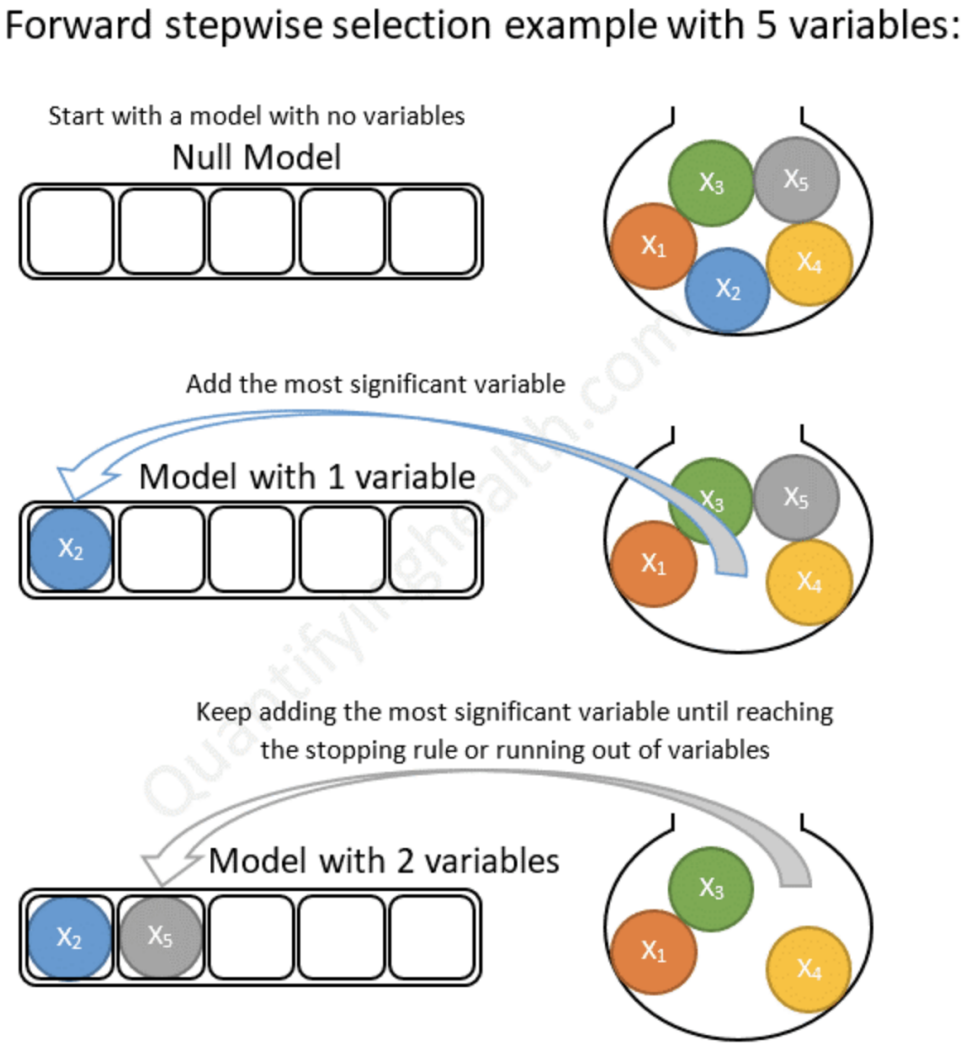
Subset Selection Methods

Best subset regression finds for each $k \in \{0, 1, 2, \cdots p\}$the subset of size k that gives smallest residual sum of squares. And then use methods such as AIC or BIC to determine the right k.

Remark

Best subset selection suffers from computational limitations. In general, there are 2^p models that involve subsets of p predictprs. If $p = 20$, there are approximately million possibilities. As for $p \geq 40$, best subset selection becomes literally infeasible.

Forward Stepwise Selection



Forward Stepwise Selection

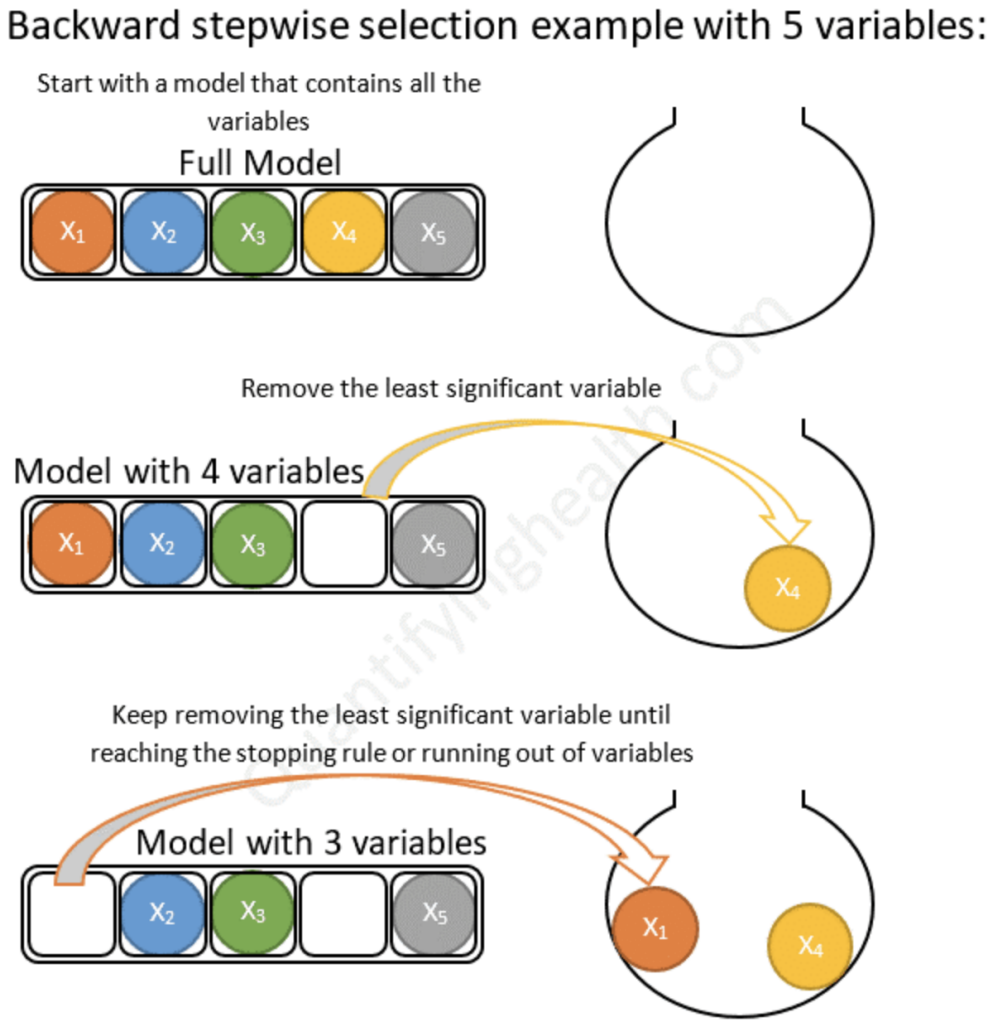
Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward Stepwise Selection

- To choose a stopping rule, we can stop when all remaining variables to consider have a p-value larger than some threshold if added to the model.
  The threshold can be:
  1. A fixed value (for instance: 0.05 or 0.2 or 0.5)
  2. Determined by AIC: AIC chooses the threshold according to how many degrees of freedom the variable under consideration has, generally speaking, the more degrees of freedom a variable has, the lower the threshold will be.
  3. Determined by BIC: BIC chooses the threshold according to the effective sample size n.

Backward Stepwise Selection



Backward Stepwise Selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Backward Stepwise Selection

- To choose a stopping rule, we can stop when all remaining variables to consider have a p-value smaller than some threshold if added to the model.
  The threshold can be:
  1. A fixed value (for instance: 0.05 or 0.2 or 0.5)
  2. Determined by AIC
  3. Determined by BIC

Comparing Backward with Forward Stepwise Selection

- Where forward stepwise is better: It can be applied in settings where the number of variables under consideration is larger than the sample size.
- Stepwise selection is more effective , but it is not guaranteed to select the best possible combination of variables. In fact the selection of variables using a stepwise regression will be highly unstable, especially when we have a small sample size.

Stagewise Selection

Algorithm 3.4 *Incremental Forward Stagewise Regression—FS_ϵ.*

1. Start with the residual $\mathbf{r}$ equal to $\mathbf{y}$ and $\beta_1, \beta_2, \ldots, \beta_p = 0$. All the predictors are standardized to have mean zero and unit norm.

2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$

3. Update $\beta_j \leftarrow \beta_j + \delta_j$, where $\delta_j = \epsilon \cdot \text{sign}[\langle \mathbf{x}_j, \mathbf{r} \rangle]$ and $\epsilon > 0$ is a small step size, and set $\mathbf{r} \leftarrow \mathbf{r} - \delta_j \mathbf{x}_j$.

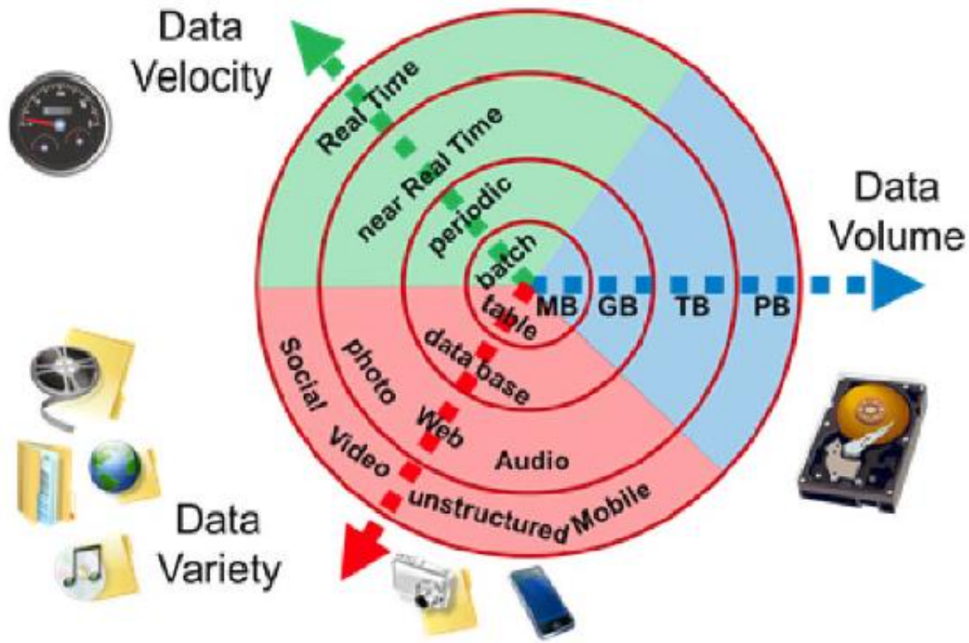
4. Repeat steps 2 and 3 many times, until the residuals are uncorrelated with all the predictors.

Comparing Stagewise with Stepwise

- Greediness is counterbalanced by the small step size $\epsilon > 0$;
- The learning process in stagewise is slowlier: In stepwise we increase the coefficient of X_j by large amount and we might change several coefficients each time; in stagewise a coefficient is increased by ϵ each time.
- The stagewise learning learning process is computationally cheap.

Introduction to Sure Independent Screening

Background: Ultrahigh dimensional data in which the number of fertures p can be much larger than the number of observations n is common but hard to deal with. Challenges includes scalability, high collinearity, spurious correlation and noise accumulation.



Introduction to Sure Independent Screening

The idea of feature screening became important for

- Computational efficiency
- Statistical efficiency: by alleviated noise accumulation through dimension reduction.

Introduction to Sure Independent Screening

Let us consider the linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

where $\boldsymbol{y}$ is an n-dimensional vector, $\boldsymbol{X}$ is an $n \times p$ matrix, $\boldsymbol{\beta}$ is a p-dimensional vector and ϵ is a p-dimensional error vector. Firstly, the SIS ranks all the p features using the marginal utilities based on the marginal correlations $\hat{corr}(\boldsymbol{x}_j, \boldsymbol{y})$ of $\boldsymbol{x}_j$s with the response of $\boldsymbol{y}$and retains the top d covariates with the largest absolute correlations collected in the set $\hat{\mathcal{M}}$, that is

$$\hat{\mathcal{M}} = \{1 \leq j \leq p : |\hat{corr}(\boldsymbol{x}_j, \boldsymbol{y})| \text{ is among the top } d \text{ largest ones}\}$$

Introduction to Sure Independent Screening

The next step of SIS: Allow user's to use any favourite regularization method such as Lasso,SCAD AND Dantzig selector on the features in $\hat{\mathcal{M}}$.

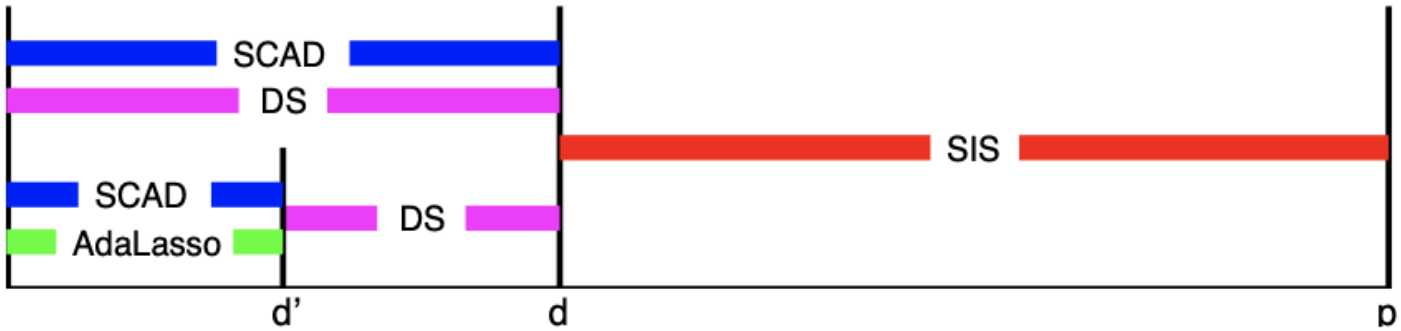


Figure 2: Methods of model selection with ultra high dimensionality.

The property this equation enjoys: $\mathbb{P}\{\boldsymbol{\mathcal{M}} \in \hat{\boldsymbol{\mathcal{M}}}\} \to 1$. It is called sure screening property and it is crucial to the 2^{nd} step of refined variable selection.

Table of Contents

Problems of Colinearity

Multicollinearity

Collinearity (or multicollinearity or ill-conditioning) occurs when independent variables in a regression are so highly correlated that it becomes difficult or impossible to distinguish their individual effects on the dependent variable.

The classical linear problem is unsolvable(i.e not having a unique solution) if there is an exact linear relationship between variables.Even if variables are highly correlated, the following problems arises:

- Confounding variables
- Computational problem
- Large variance

Confounding Variables

For simplicity, let's assume complete linear relationship between variables. In the following example, we assume $X_1 = X_2$, and try to fit the following mode(Note that when we do regression with X_1 only we get $Y = 20 - 3X_1$) :

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2,$$

all the above solutions are equivalent

$$Y = 20 - 5X_1 + 2X_2$$

$$Y = 20 + 6X_1 - 3X_2$$

$$Y = 20 + 3X_2$$

But the meaning of each equation can be significantly different since the coefficients of each one implies a different relation

Computational Problem

Definition of condition number: Let e be the error in b. Assuming that A is a nonsingular matrix, the error in the solution of $AX = b$ $A^{-1}b$ is $A^{-1}e$. The maximum value of the ratio of the relative error

$$\max_{e,b\neq 0}\left\{\frac{\|A^{-1}e\|}{\|e\|}\frac{\|b\|}{\|A^{-1}b\|}\right\} = \max_{e\neq 0}\left\{\frac{\|A^{-1}e\|}{\|e\|}\right\}\max_{b\neq 0}\left\{\frac{\|b\|}{\|A^{-1}b\|}\right\}$$
$$= \|A^{-1}\|\,\|A\|$$

Take$\|\cdot\|$ to be $\|\cdot\|_2$, then:

$$\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$$

Remark

The matrix X^TX is 'close' to singular. This means its smallest eigenvalue is close to zero, thus its condition number is large and the solution is sensitive to disturbance.

Large Variance

We have learned previously that $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \left(\mathrm{X}^T\mathrm{X}\right)^{-1}\sigma^2$, with

$$\mathrm{Var}\left(\hat{\beta}_p\right) = \frac{\sigma^2}{\langle \mathrm{z}_p, \mathrm{z}_p \rangle} = \frac{\sigma^2}{\|\mathrm{z}_p\|^2}$$

being the variance of a individual coefficient(z_p being the variable we get from Orthogonalization).
For collinearity problem, if two columns in $\boldsymbol{X}$, say X_i, X_j is highly correlated, then the length of z_j must be rather small as most of the information in z_j has already been explained by z_i, making the variance very large.

Examples of Colinearity

In reality, we give an example of generating highly correlated variables:

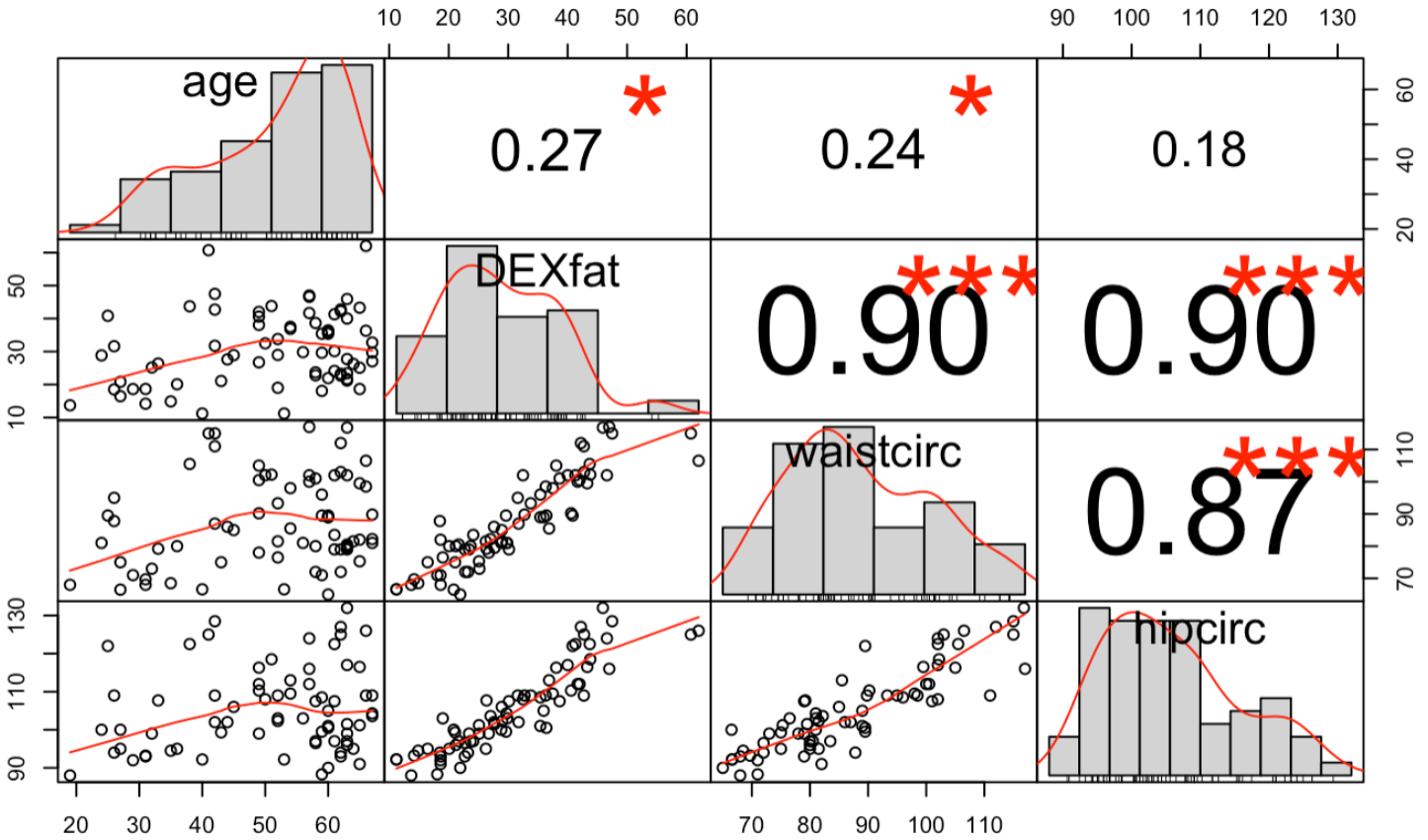
- Suppose(x, y) is generated as follows:

$$
\begin{aligned}
z_1, z_2 &\sim \mathcal{N}(0,1) \text{ (independent)} \\
\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_6 &\sim \mathcal{N}(0,1) \text{ (independent)} \\
y &= 3z_1 - 1.5z_2 + 2\varepsilon_0 \\
x_j &= \begin{cases} z_1 + \varepsilon_j/5 & \text{for } j = 1,2,3 \\ z_2 + \varepsilon_j/5 & \text{for } j = 4,5,6 \end{cases}
\end{aligned}
$$

- Generated a sample of (x, y) pairs of size 100.
- Correlations within the groups of x's were around 0.97.

Examples of Colinearity

The following is a dataset called 'bodyfat'.



DEXfat:body fat measured by DXA, response variable;
waistcirc:waist circumference; hipcirc: hip circumference.

Shrinkage Method

The idea here is to perform a linear regression, while shrinking the regression coefficients $\hat{\beta}$ toward 0.

- Wildly large positive/negative coefficients on one variables became less likely to occur.
- Discrete process(i.e variables are either retained or discarded) often exhibits high variance.
- Shrinkage methods are more continuous and don't suffer from high variability.
- It can reduce high variance of regression coefficient caused by collinearity between explanatory variables.

Ridge Regression Model

Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares,

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

An equivalent way to write the ridge problem is

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq t$$

Ridge Regression Solution

Writing in matrix form,

$$\mathrm{RSS}(\lambda) = (\mathrm{y} - \mathrm{X}\beta)^T(\mathrm{y} - \mathrm{X}\beta) + \lambda\beta^T\beta$$

The regression solution is

$$\hat{\beta}^{\mathrm{ridge}} = \left(\mathrm{X}^T\mathrm{X} + \lambda\mathrm{I}\right)^{-1}\mathrm{X}^T\mathrm{y}$$

Proof:let

$$X_* = \begin{pmatrix} X \\ vI \end{pmatrix}$$

Solving ridge estimator for X is equivalent to solving OLS for X_*

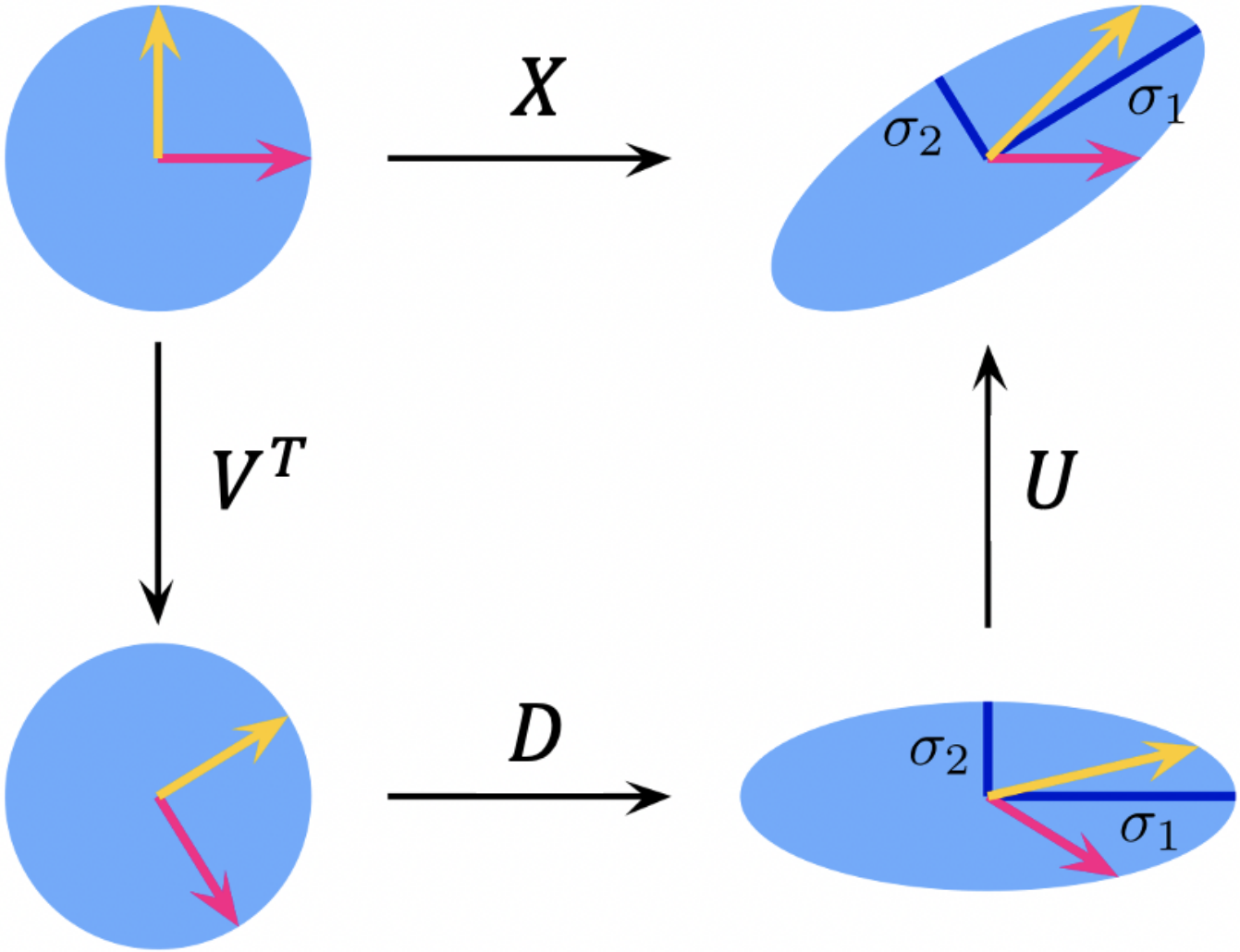
$$(y_* - X_*\beta)'(y_* - X_*\beta) = (y - X\beta)'(y - X\beta) + \lambda\beta'\beta$$

$$\left(X_*^T X_*\right)\beta = X_*' y_*$$

Plug in X_*, and we get

$$\left(X^T X + \lambda I\right)\beta = X^T y$$

SVD Decmoposition



$$X = UDV^T$$

Bias of Ridge Estimator

Under the commonly assumption about the noise:

$$\mathrm{E}[\varepsilon \mid \mathrm{X}] = 0$$

$$\mathrm{Var}[\varepsilon \mid X] = \sigma^2 I$$

The bias of $X\widehat{\beta}_\lambda$ satisfies:

$$Bias^2(X\widehat{\beta}_\lambda) = \|X(\mathbb{E}_\varepsilon \beta_\lambda - \beta)\| \leq \lambda \|\beta\|_2$$

First consider $\mathbb{E}_\varepsilon \beta$ is the minimizer of $\|X\beta - X\gamma\|^2 + \lambda\|\gamma\|_2$

$$\begin{aligned} \|X(\mathbb{E}_\varepsilon \beta_\lambda - \beta)\| &\leq \|X(\mathbb{E}_\varepsilon \beta_\lambda - \beta)\| + \lambda\|\hat{\beta}_\lambda\|_2 \\ &\leq \|X\beta - X\gamma\|^2 + \lambda\|\gamma\|_2. \end{aligned}$$

This suggests that the solution of ridge regression is actually a biased estimator of β and bias increase with λ.

Variance of Ridge Estimator

$\text{Var}\left[X\widehat{\beta}_\lambda\right] = \sigma^2 \sum_{i=1}^n \mu_i^2$,where $\mu_i = \frac{\sigma_i^2}{\sigma_i^2+\lambda}$

Proof:

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}_\varepsilon \left\| X\left(\mathbb{E}\hat{\beta}_\lambda - \hat{\beta}_\lambda\right)\right\|_2^2 \\
=&\frac{1}{n}\mathbb{E}_\varepsilon \| X\left(X^\top X + \lambda I\right)^{-1} X^\top X\beta - X\left(X^\top X + \lambda I\right)^{-1} X^\top (X\beta + \varepsilon)\|^2 \\
=&\frac{1}{n}\mathbb{E}_\varepsilon \| W\varepsilon\|^2 = \frac{\sigma^2}{n}\sum_{i=1}^n \mu_i^2,
\end{aligned}$$

where μ_i is the eigenvalues of W, and $W = X(X^TX + \lambda I)^{-1}X^T$. Let $X = UDV^T$, and

$$W = UDV^T\left(VD^2V^T + \lambda I\right)^{-1} = \quad UD(D^2 + \lambda I)^{-1}DU^T,$$

thus $\mu_i = \frac{\sigma_i^2}{\sigma_i^2+\lambda}$, which suggests $\text{Var}\left[X\widehat{\beta}_\lambda\right] < \text{Var}\left[X\widehat{\beta}\right]$.

SVD Decomposition

The singular value decomposition (SVD) of the centered input matrix X is

$$\mathrm{X} = \mathrm{U}_{\mathrm{n}\times\mathrm{p}}\mathrm{D}_{\mathrm{p}\times\mathrm{p}}\mathrm{V}_{\mathrm{p}\times\mathrm{p}}{}^{T}.$$

Here U and V are orthogonal matrices, and D is a diagonal matrix, with diagonal entries $d_1 \geq d_2 \geq \cdots d_p \geq 0$. Using the SVD of X we have:

$$\begin{aligned}
\mathrm{X}\hat{\beta}^{\mathrm{1s}} &= \mathrm{X}\left(\mathrm{X}^{T}\mathrm{X}\right)^{-1}\mathrm{X}^{T}\mathrm{y} \\
&= \mathrm{U}\mathrm{U}^{T}\mathrm{y} = \sum_{j=1}^{n}\mathrm{u}\mathrm{u}_{\mathrm{j}}^{\mathrm{T}}y \\
\mathrm{X}\hat{\beta}^{\mathrm{ridge}} &= \mathrm{X}\left(\mathrm{X}^{T}\mathrm{X}+\lambda\mathrm{I}\right)^{-1}\mathrm{X}^{T}\mathrm{y} \\
&= \mathrm{U}\mathrm{D}\left(\mathrm{D}^{2}+\lambda\mathrm{I}\right)^{-1}\mathrm{D}\mathrm{U}^{T}\mathrm{y} \\
&= \sum_{j=1}^{p}\mathrm{u}_{j}\frac{d_j^2}{d_j^2+\lambda}\mathrm{u}_{j}^{T}\mathrm{y}
\end{aligned}$$

SVD Decomposition

Note that for $\lambda \geq 0$, we have $\frac{d_j^2}{d_j^2+\lambda} \leq 1$

- Like linear regression,ridge regression computes the coordinates of y w.r.t the orthonormal basis U.
- It shrinks these coordinates by the factor $\frac{d_j^2}{d_j^2+\lambda}$.
- A greater amount of shrinkage is applied to the coordinates of basis vectors with smaller d_j^2.
- We define the effective degrees of freedom as follows:

$$\mathrm{df}(\lambda) = \sum_{j=1}^{D} \frac{d_j^2}{d_j^2 + \lambda}$$

Note that $df(\lambda) = p$ when $\lambda = 0$(no regularization);
and $df(\lambda) \to 0$ as $\lambda \to \infty$.

PCA as a Variance-Maximization Technique

Principal component analysis (PCA) was introduced as a technique for deriving a reduced set of orthogonal linear projections of a single collection of correlated variables, $\mathrm{X} = (X_1, \ldots, X_p)$, where the projections are odered by decreasing variances.

Variance is an important measurement of the amount of information in the variable.

PCA can be viewed as a optimization problem as following:

$$\max \mathrm{Var}\,(Xb_j)$$

$$s.t\ \ b_j^T b_j = 1, \mathrm{Var}\,(Xb_i, Xb_j) = b_i \Sigma_{XX} b_j = 0, i < j,$$

$$notice\ \ that\ (Xb_i, Xb_j) = b_i \Sigma_{XX} b_j$$

PCA as a Variance-Maximization Technique

The objective function is:

$$f(\mathrm{b}_1) = \mathrm{b}_1^T \Sigma_{XX} \mathrm{b}_1 - \lambda_1 \left(1 - \mathrm{b}_1^\top \mathrm{b}_1\right),$$

λ_1 is a Lagrange multiplier.Differentiating $f(\mathrm{b}_1)$ with respect to b_1 and setting the result equal to zero for a maximum to yield.

$$\frac{\partial f(\mathrm{b}_1)}{\partial \mathrm{b}_1} = 2\left(\mathbf{\Sigma}_{XX} - \lambda_1 \mathrm{I}_r\right)\mathrm{b}_1 = 0$$

If $\mathrm{b}_1 \neq 0$, then λ_1 must be chosen to satisfy the determinantal equation

$$|\mathbf{\Sigma}_{XX} - \lambda_1| = 0$$

Thus, λ_1 has to be the largest eigenvalue of $\mathbf{\Sigma}_{XX}$, and b_1 be the eigenvector associated with λ_1.

PCA as a Variance-Maximization Technique

The second principal component, ξ_2, is then obtained by choosing a second set of coefficients, b_2, for the next linear projection, ξ_2, so that the variance of ξ_2 is largest among all linear projections of X that are also uncorrelated with ξ_1 above.
The variance of ξ_2 is $\mathrm{Var}(\xi_2) = b_2^T \Sigma_{XX} b_2$, and this has to be maximized subject to the normalization constraint $b_2^T b_2 = 1$ and orthogonality constraint $b_1^T b_2 = 0$.

$$f(b_2) = b_2^T \Sigma_{XX} b_2 - \lambda_2 \left(1 - b_2^T b_2\right) + \mu b_1^T b_2$$

Differentiating $f(b_2)$ with respect to b_2 and setting the result equal to zero for a maximum yields

$$\frac{\partial f(b_2)}{\partial b_2} = 2\left(\Sigma_{XX} - \lambda_2 I_r\right) b_2 + \mu b_1 = 0$$

PCA as a Variance-Maximization Technique

Premultiplying this derivative by $\mathbf{b}_1^T$ and using the orthogonality and normalization constraints, we have that

$$2\mathbf{b}_1^T \boldsymbol{\Sigma}_{XX} \mathbf{b}_2 + \mu = 0.$$

Premultiplying the equation $(\boldsymbol{\Sigma}_{XX} - \lambda_1 \mathbf{I}_r)\,\mathbf{b}_1 = 0$ by $\mathbf{b}_2^T$ yields

$$\mathbf{b}_2^T \boldsymbol{\Sigma}_{XX} \mathbf{b}_1 = 0$$

whence $\mu = 0$.
This means that λ_2 is the second largest eigenvalue of $\boldsymbol{\Sigma}_{XX}$, and the coefficient vector $\mathbf{b}_2$ for the second principal component is the eigenvector associated with λ_2.

PCA as a Variance-Maximization Technique

In this sequential manner, we obtain the remaining sets of coefficients for the principal components $\xi_3, \xi_4, \ldots, \xi_r$, where the ith principal component ξ_i is obtained by choosing the set of coefficients, b_i, for the linear projection ξ_i so that ξ_i has the largest variance among all linear projections of X that are also uncorrelated with $\xi_1, \xi_2, \ldots, \xi_{i-1}$.
The coefficients of these linear projections are given by the ordered sequence of eigenvectors $\{\mathrm{b}_j\}$ associated with the jth largest eigenvalue λ_j, of $\mathbf{\Sigma}_{XX}$.

Principe Components of X

We use the SVD of the centered matrix X to express the principal components of the variables in X. The sample covariance matrix is given by

$$\mathrm{S} = \mathrm{X}^T\mathrm{X}/(N-1).$$

And the SVD of S is

$$S = \mathrm{X}^T\mathrm{X}/(N-1) = \mathrm{V}\mathrm{D}^2\mathrm{V}^T/(N-1).$$

The first eigenvectors v_j has the property that: $z_1 = Xv_1$ has the largerst variance among all $z_j = Xv_j$:

$$\mathrm{Var}\,(\mathrm{z}_1) = \mathrm{Var}\,(\mathrm{X}v_1) = \frac{d_1^2}{N-1},$$

In fact, $z_1 = Xv_1 = u_1 d_1$, hence u_1 is the nomalized first PC.

PCA with Least-Square Optimality

Let B be a $(t \times p)$ matix $\mathrm{B} = (\mathrm{b}_1, \cdots, \mathrm{b}_t)^\tau$. Let $\xi = \mathrm{BX}$, we want to find a p-vector μ $(p \times t)$ matix A.

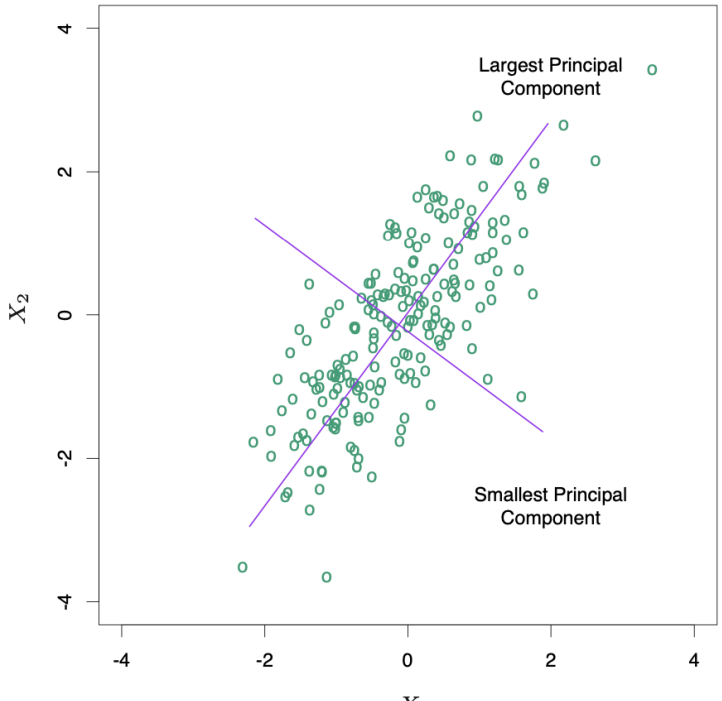
$$\mathrm{E}\left\{(\mathrm{X} - \boldsymbol{\mu} - \mathrm{A}\xi)^T(\mathrm{X} - \boldsymbol{\mu} - \mathrm{A}\xi)\right\},$$

we use least-squares error criterion as our measure of how well we can reconstruct X by linear projection on ξ.
More specifically, our goal is to choose A, B and μ to minimize

$$E\left\{(\mathrm{X} - \boldsymbol{\mu} - \mathrm{ABX})^T(\mathrm{X} - \boldsymbol{\mu} - \mathrm{ABX})\right\}.$$

PCR

The PCR method may be broadly divided into three major steps:
1.Obtain the principal components following the above step;
2. Regress the observed vector of outcomes on the selected first k principal components as covariates, get OLS.(notice we can take different value of k)
3. Decide for the best k(Here we present one rule: We find the least k that satisfies: $\sum d_k \geq 0.8$)



Compare PCR with Ridge Regression

Let $W_k = \mathrm{X}V_k = [\mathrm{X}\mathrm{v}_1, \ldots, \mathrm{X}\mathrm{v}_k]$, $\widehat{\gamma}_k = \left(W_k^T W_k\right)^{-1} W_k^T \mathrm{Y} \in \mathbb{R}^k$ we here give the solution of PCR with k principle components:

$$X\widehat{\beta}_k = V_k\widehat{\gamma}_k = \sum_{j=1}^{k} u_j u_j^T y$$

It is similar to ridge regression in the way that it also reduced variance of the coefficients:

$$\begin{aligned}
\mathrm{Var}\left(\widehat{\beta}_k\right) &= \sigma^2 V_k \left(W_k^T W_k\right)^{-1} V_k^T \\
&= \sigma^2 V_k \,\mathrm{diag}\left(\lambda_1^{-1}, \ldots, \lambda_k^{-1}\right) V_k^T \\
&= \sigma^2 \sum_{j=1}^{k} \frac{\mathrm{v}_j \mathrm{v}_j^T}{\lambda_j}
\end{aligned}$$

$$\mathrm{Var}\left(\widehat{\beta}_{\mathrm{ols}}\right) - \mathrm{Var}\left(\widehat{\beta}_k\right) = \sigma^2 \sum_{j=k+1}^{p} \frac{\mathrm{v}_j \mathrm{v}_j^T}{\lambda_j}$$

Compare PCR with Ridge Regression

Also, multicollinearity can be addressed by both PCR and ridge regression. We have known from the above analysis of the variance of small estimator that it is the small eigenvalue that has the maximum inflation effect on the variance of the least squares estimator.
In fact,PCR eliminates the effect on those eigenvectors by simply exclude them from the regression model$\sum_{j=1}^{k} u_j u_j^T y$.
On the other hand,ridge regression eliminates the effect by multiplying a coefficient $\frac{d_j^2}{d_j^2+\lambda}$for each term $\sum_{j=1}^{p} \mathrm{u}_j \frac{d_j^2}{d_j^2+\lambda} \mathrm{u}_j^T \mathrm{y}$. Ridge regression projects y onto every principle components, but shrinks the coefficients of the low variance components more than high ones since lower ones contains less information.

Table of Contents

Appendix

(Cochran's Theorem) Let $X_1, X_2, \ldots, X_n$ be independent $N\left(\mu_i, \sigma^2\right)$, i.e $\mathrm{X} \sim \mathrm{N}\left(\boldsymbol{\mu}, \sigma^2 \mathrm{I}\right)$, and

$$\sum_{i=1}^{n} X_i^2 = Q_1 + Q_2 + \cdots + Q_k,$$

where $Q_i = \mathrm{X}'\mathrm{A}_i\mathrm{X}, \mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_k$ are symmetric and idempotent $n \times n$ matrices with $\operatorname{rank}(\mathrm{A}_i) = r_i, i = 1, 2, \ldots, k$.
Then:
(1) $Q, Q_2, \ldots, Q_k$ are independent.
(2) $\frac{Q_i}{\sigma^2} \sim \chi^2\left(r_i, \delta_i\right)$ with $\delta_i = \boldsymbol{\mu}'\mathrm{A}_i\boldsymbol{\mu}'/\sigma^2$.

Corollary: Let $\mathrm{X} \sim \mathrm{N}\left(\boldsymbol{\mu}, \sigma^2 \mathrm{I}\right)$, A is symmetric with $\operatorname{rank}(\mathrm{A}) = r$, $\delta = \boldsymbol{\mu}'\mathrm{A}\boldsymbol{\mu}'/\sigma^2$. Then

$$\mathrm{X}'\mathrm{AX}/\sigma^2 \sim \chi^2(r, \delta) \Leftrightarrow \mathrm{A} \text{ is idempotent .}$$

Appendix

(Cook's Distance)

$$\hat{\boldsymbol{\beta}}_{(i)} = \left(\mathbf{X}_{(i)}^{\mathrm{T}}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^{\mathrm{T}}\mathbf{y}_{(i)},$$

$$\mathbf{X}_{(i)}^{\mathrm{T}}\mathbf{X}_{(i)} = \mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}},$$

$$\left(\mathbf{X}_{(i)}^{\mathrm{T}}\mathbf{X}_{(i)}\right)^{-1} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} + \frac{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}}{1 - \mathbf{x}_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}_i}.$$

$$\hat{\mathbf{y}}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)},$$

$$\hat{\boldsymbol{\beta}}_{(i)} = \left(\mathbf{X}_{(i)}^{\mathrm{T}}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^{\mathrm{T}}\mathbf{y}_{(i)},$$

$$h_i = H_{ii} = \mathbf{x}_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}_i,$$

$$\hat{\boldsymbol{\beta}}_{(i)} = \left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} + \frac{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}}{1 - h_i}\right)\left(\mathbf{X}^{\mathrm{T}}\mathbf{y} - y_i\mathbf{x}_i\right).$$

Appendix

(Cook's Distance)

$$\hat{\boldsymbol{\beta}} = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y},$$

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} + \frac{\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{x}_i\mathrm{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{1-h_i} - y_i\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{x}_i - \frac{h_i y_i}{1-h_i}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{x}_i.$$

$$y_i = \mathrm{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{y_i - \hat{y}_i}{1-h_i}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{x}_i = \hat{\boldsymbol{\beta}} - \frac{\hat{\epsilon}_i}{1-h_i}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{x}_i,$$

$$y_i - \hat{y}_{(i)} = y_i - \mathrm{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(i)}\hat{\boldsymbol{\beta}}_{(i)},$$

$$y_i - \hat{y}_{(i)} = y_i - \mathrm{x}_i^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}} - \frac{\hat{\epsilon}_i}{1-h_i}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{x}_i\right),$$

$$y_i - \hat{y}_{(i)} = \hat{\epsilon}_i + \frac{\hat{\epsilon}_i h_i}{1-h_i} = \frac{\hat{\epsilon}_i}{1-h_i}.$$