Machine Learning I

Linear Regression

Souhaib Ben Taieb

University of Mons

UMONS Université de Mons

Faculté des Sciences

Table of contents

Simple linear regression

- Optimal predictions
- Parameter estimation with least squares
- Parameter estimation with MLE
- Model accuracy and hypothesis testing

Multiple linear regression

- Parameter estimation
- Interpreting regression coefficients
- Qualitative/categorical variables
- Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

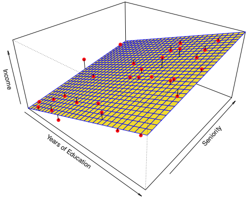
- How to estimate the out-of-sample error in linear regression?
- Variable subset selection methods
- Shrinkage methods

Linear regression models

- Although optimal prediction functions are *very rarely linear*, **linear models** are useful both conceptually and practically.
- The **squared error loss function** is often used in (linear) regression, i.e. $L(y, h(x)) = (y - h(x))^2$.
- The **linear hypothesis set** is composed of affine (linear) functions, i.e.

$$\mathcal{H}_{\text{lin}} = \{h(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p | \beta_0, \beta_1, \ldots, \beta_p \in \mathbb{R}\},$$

where $x = [x_1, x_2, \ldots, x_p]^T \in \mathbb{R}^p$.



Assumptions

When fitting a linear regression model to data, we try to find the best prediction function in a **linear hypothesis set** but

- We **do not** assume that the relationship between x and y really is linear.
- We **do not** assume anything about the marginal distributions of x and y, or about their joint distributions.

Outline

Simple linear regression

Optimal predictions

Parameter estimation with least squares

Parameter estimation with MLE

Model accuracy and hypothesis testing

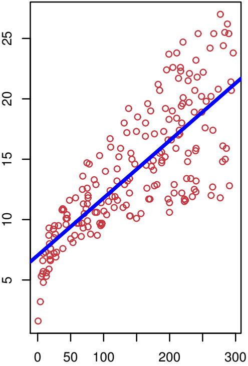
Multiple linear regression

Linear Model Selection and Regularization

Simple linear regression

We will first consider *simple linear regression*, i.e. linear regression with a single input $x \in \mathbb{R}$ ($p = 1$) with the hypothesis set

$$\mathcal{H} = \{h(x) = \beta_0 + \beta_1 x | \beta_0, \beta_1 \in \mathbb{R}\},$$



Outline

Simple linear regression

- Optimal predictions
- Parameter estimation with least squares
- Parameter estimation with MLE
- Model accuracy and hypothesis testing

Multiple linear regression

- Parameter estimation
- Interpreting regression coefficients
- Qualitative/categorical variables
- Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

- How to estimate the out-of-sample error in linear regression?
- Variable subset selection methods
- Shrinkage methods

Optimal predictions

What are the **optimal predictions** in simple linear regression? In other words, we want to compute

$$g^* = \underset{h \in \mathcal{H}_{\text{lin}}}{\operatorname{argmin}} \underbrace{\mathbb{E}_{x,y}[(y - h(x))^2]}_{E_{\text{out}}(h)},$$

or, equivalently,

$$(\beta_0^*, \beta_1^*) = \underset{(\beta_0,\beta_1) \in \mathbb{R}^2}{\operatorname{argmin}} \underbrace{\mathbb{E}_{x,y}[(y - (\beta_0 + \beta_1 x))^2]}_{E_{\text{out}}(\beta_0,\beta_1)},$$

since β_0 and β_1 completlety characterize $h(x) = \beta_0 + \beta_1 x$.

We can show that the **optimal coefficients** are

$$\boxed{\beta_1^* = \frac{\operatorname{Cov}(x, y)}{\operatorname{Var}(x)}} \text{ and } \boxed{\beta_0^* = \mathbb{E}[y] - \beta_1^* \mathbb{E}[x]}.$$

Optimal predictions

What are the **optimal predictions** in simple linear regression? In other words, we want to compute

$$g^* = \underset{h \in \mathcal{H}_{\text{lin}}}{\operatorname{argmin}} \underbrace{\mathbb{E}_{x,y}[(y - h(x))^2]}_{E_{\text{out}}(h)},$$

or, equivalently,

$$(\beta_0^*, \beta_1^*) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\operatorname{argmin}} \underbrace{\mathbb{E}_{x,y}[(y - (\beta_0 + \beta_1 x))^2]}_{E_{\text{out}}(\beta_0, \beta_1)},$$

since β_0 and β_1 completlety characterize $h(x) = \beta_0 + \beta_1 x$.

We can show that the **optimal coefficients** are

$$\boxed{\beta_1^* = \frac{\text{Cov}(x, y)}{\text{Var}(x)}} \text{ and } \boxed{\beta_0^* = \mathbb{E}[y] - \beta_1^* \mathbb{E}[x]}.$$

Optimal predictions

What are the **optimal predictions** in simple linear regression? In other words, we want to compute

$$g^* = \underset{h \in \mathcal{H}_{\text{lin}}}{\operatorname{argmin}} \underbrace{\mathbb{E}_{x,y}[(y - h(x))^2]}_{E_{\text{out}}(h)},$$

or, equivalently,

$$(\beta_0^*, \beta_1^*) = \underset{(\beta_0,\beta_1) \in \mathbb{R}^2}{\operatorname{argmin}} \underbrace{\mathbb{E}_{x,y}[(y - (\beta_0 + \beta_1 x))^2]}_{E_{\text{out}}(\beta_0,\beta_1)},$$

since β_0 and β_1 completlety characterize $h(x) = \beta_0 + \beta_1 x$.

We can show that the **optimal coefficients** are

$$\boxed{\beta_1^* = \frac{\text{Cov}(x, y)}{\text{Var}(x)}} \text{ and } \boxed{\beta_0^* = \mathbb{E}[y] - \beta_1^* \mathbb{E}[x]}.$$

Outline

Simple linear regression

Optimal predictions

Parameter estimation with least squares

Parameter estimation with MLE

Model accuracy and hypothesis testing

Multiple linear regression

Parameter estimation

Interpreting regression coefficients

Qualitative/categorical variables

Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

How to estimate the out-of-sample error in linear regression?

Variable subset selection methods

Shrinkage methods

Parameter estimation (model fitting)

Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ where $x_i, y_i \in \mathbb{R}$, we compute

$$g = \underset{h \in \mathcal{H}_{\text{lin}}}{\operatorname{argmin}} \ \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - h(x_i))^2}_{E_{\text{in}}(h)},$$

or, equivalently,

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1) \in \mathbb{R}^2}{\operatorname{argmin}} \ \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2}_{E_{\text{in}}(\beta_0, \beta_1)}.$$

The **solution** can be shown to be

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

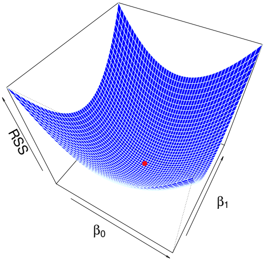
where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

The least squares method

If we let $e_i = y_i - (\beta_0 + \beta_1 x_i)$ represent the ith **residual**, we define the **residual sum of squares** (RSS) as

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2.$$

Minimizing $E_{\text{in}} = \frac{\mathrm{RSS}}{n}$. is equivalent to minimize RSS, which is known as the **(ordinary) least squares (OLS)** method.



Bias and variance of $\hat{\beta}_1$ and $\hat{\beta}_0$

Let us **assume** the *data generating process* is given by:

$$y = \beta_0^* + \beta_1^* x + \varepsilon, \tag{1}$$

where β_0^* and β_1^* are the true coefficients, ε is a random noise term with $\mathbb{E}[\varepsilon|x] = 0$ and $\text{Var}(\varepsilon|x) = \sigma^2$.

Then we can show that

$$\mathbb{E}[\hat{\beta}_1] = \beta_1^* \quad \text{and} \quad \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

and

$$\mathbb{E}[\hat{\beta}_0] = \beta_0^* \quad \text{and} \quad \text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right].$$

Bias and variance of $\hat{\beta}_1$ and $\hat{\beta}_0$

Let us **assume** the *data generating process* is given by:

$$y = \beta_0^* + \beta_1^* x + \varepsilon, \tag{1}$$

where β_0^* and β_1^* are the true coefficients, ε is a random noise term with $\mathbb{E}[\varepsilon|x] = 0$ and $\mathrm{Var}(\varepsilon|x) = \sigma^2$.

Then we can show that

$$\mathbb{E}[\hat{\beta}_1] = \beta_1^* \quad \text{and} \quad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

and

$$\mathbb{E}[\hat{\beta}_0] = \beta_0^* \quad \text{and} \quad \mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

Summary

$$g^*(x) = \beta_0^* + \beta_1^* x$$

$$\beta_1^* = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

$$\beta_0^* = \mathbb{E}[y] - \beta_1^* \mathbb{E}[x].$$

$$g(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

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

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

$$\mathbb{E}[\hat{\beta}_1] = \beta_1^*$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\mathbb{E}[\hat{\beta}_0] = \beta_0^*$$

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right],$$

Replacing the population quantities with their sample counterparts is known as the "**plug-in principle**".

Outline

Simple linear regression

Optimal predictions

Parameter estimation with least squares

Parameter estimation with MLE

Model accuracy and hypothesis testing

Multiple linear regression

Parameter estimation

Interpreting regression coefficients

Qualitative/categorical variables

Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

How to estimate the out-of-sample error in linear regression?

Variable subset selection methods

Shrinkage methods

Maximum Likelihood Estimation

Given a parametric distribution under consideration, the goal of maximum likelihood estimation (MLE) is to select the distribution that is **most likely** to have generated the sample $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$.

In simple linear regression, we often consider the distribution $p_{y|x}(y|x; \boldsymbol{\theta}) = \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$ where $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma)$, $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma > 0$. This is equivalent to assume

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{2}$$

where $\varepsilon | x \sim \mathcal{N}(0, \sigma^2)$, i.e. ϵ is independent of x.

If the dataset $\mathcal{D}$ has a DGP given by (2) with **i.i.d**. ε_i, y_i and y_j are independent given x_i and x_j $(i \neq j)$. Let us compute the **(conditional) (log-)likelihood function**.

Maximum Likelihood Estimation

Given a parametric distribution under consideration, the goal of maximum likelihood estimation (MLE) is to select the distribution that is **most likely** to have generated the sample $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

In simple linear regression, we often consider the distribution $p_{y|x}(y|x; \boldsymbol{\theta}) = \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$ where $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma)$, $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma > 0$. This is equivalent to assume

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{2}$$

where $\varepsilon | x \sim \mathcal{N}(0, \sigma^2)$, i.e. ϵ is independent of x.

If the dataset $\mathcal{D}$ has a DGP given by (2) with **i.i.d**. ε_i, y_i and y_j are independent given x_i and x_j $(i \neq j)$. Let us compute the **(conditional) (log-)likelihood function**.

Maximum Likelihood Estimation

Given a parametric distribution under consideration, the goal of maximum likelihood estimation (MLE) is to select the distribution that is **most likely** to have generated the sample $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$.

In simple linear regression, we often consider the distribution $p_{y|x}(y|x; \boldsymbol{\theta}) = \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$ where $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma)$, $\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma > 0$. This is equivalent to assume

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{2}$$

where $\varepsilon | x \sim \mathcal{N}(0, \sigma^2)$, i.e. ϵ is independent of x.

If the dataset $\mathcal{D}$ has a DGP given by (2) with **i.i.d**. ε_i, y_i and y_j are independent given x_i and x_j $(i \neq j)$. Let us compute the **(conditional) (log-)likelihood function**.

Maximum Likelihood Estimation

Recall that if $y|x \sim \mathcal{N}(h(x), \sigma^2)$ where $h : \mathbb{R} \to \mathbb{R}$, then the conditional PDF is given by

$$p(y|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-h(x)}{\sigma}\right)^2}.$$

In simple linear regression, recall that $h(x) = \beta_0 + \beta_1 x$. The **(conditional) likelihood function** is given by

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1, \sigma) &\equiv \mathcal{L}(\beta_0, \beta_1, \sigma; \mathcal{D}) \\
&= p(y_1, \dots, y_n | x_1, \dots, x_n; \beta_0, \beta_1, \sigma) \\
&= \Pi_{i=1}^{n} p(y_i | x_1, \dots, x_n; \beta_0, \beta_1, \sigma) \\
&= \Pi_{i=1}^{n} p_{y|x}(y_i | x_i; \beta_0, \beta_1, \sigma) \\
&\propto \sigma^{-n} exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}.
\end{aligned}$$

The **(conditional) log-likelihood** is given by

$$\log \mathcal{L}(\beta_0, \beta_1, \sigma) \propto -nlog(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Maximum Likelihood Estimation

Recall that if $y|x \sim \mathcal{N}(h(x), \sigma^2)$ where $h : \mathbb{R} \to \mathbb{R}$, then the conditional PDF is given by

$$p(y|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-h(x)}{\sigma}\right)^2}.$$

In simple linear regression, recall that $h(x) = \beta_0 + \beta_1 x$. The **(conditional) likelihood function** is given by

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1, \sigma) &\equiv \mathcal{L}(\beta_0, \beta_1, \sigma; \mathcal{D}) \\
&= p(y_1, \ldots, y_n | x_1, \ldots, x_n; \beta_0, \beta_1, \sigma) \\
&= \Pi_{i=1}^{n} p(y_i | x_1, \ldots, x_n; \beta_0, \beta_1, \sigma) \\
&= \Pi_{i=1}^{n} p_{y|x}(y_i | x_i; \beta_0, \beta_1, \sigma) \\
&\propto \sigma^{-n} exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}.
\end{aligned}$$

The **(conditional) log-likelihood** is given by

$$\log \mathcal{L}(\beta_0, \beta_1, \sigma) \propto -nlog(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Maximum Likelihood Estimation

Recall that if $y|x \sim \mathcal{N}(h(x), \sigma^2)$ where $h : \mathbb{R} \to \mathbb{R}$, then the conditional PDF is given by

$$p(y|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-h(x)}{\sigma}\right)^2}.$$

In simple linear regression, recall that $h(x) = \beta_0 + \beta_1 x$. The **(conditional) likelihood function** is given by

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1, \sigma) &\equiv \mathcal{L}(\beta_0, \beta_1, \sigma; \mathcal{D}) \\
&= p(y_1, \ldots, y_n | x_1, \ldots, x_n; \beta_0, \beta_1, \sigma) \\
&= \Pi_{i=1}^n p(y_i | x_1, \ldots, x_n; \beta_0, \beta_1, \sigma) \\
&= \Pi_{i=1}^n p_{y|x}(y_i | x_i; \beta_0, \beta_1, \sigma) \\
&\propto \sigma^{-n} exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \right\}.
\end{aligned}$$

The **(conditional) log-likelihood** is given by

$$\log \mathcal{L}(\beta_0, \beta_1, \sigma) \propto -nlog(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Maximum Likelihood Estimation

Recall that if $y|x \sim \mathcal{N}(h(x), \sigma^2)$ where $h : \mathbb{R} \to \mathbb{R}$, then the conditional PDF is given by

$$p(y|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-h(x)}{\sigma}\right)^2}.$$

In simple linear regression, recall that $h(x) = \beta_0 + \beta_1 x$. The **(conditional) likelihood function** is given by

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1, \sigma) &\equiv \mathcal{L}(\beta_0, \beta_1, \sigma; \mathcal{D}) \\
&= p(y_1, \ldots, y_n | x_1, \ldots, x_n; \beta_0, \beta_1, \sigma) \\
&= \Pi_{i=1}^{n} p(y_i | x_1, \ldots, x_n; \beta_0, \beta_1, \sigma) \\
&= \Pi_{i=1}^{n} p_{y|x}(y_i | x_i; \beta_0, \beta_1, \sigma) \\
&\propto \sigma^{-n} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}.
\end{aligned}$$

The **(conditional) log-likelihood** is given by

$$\log \mathcal{L}(\beta_0, \beta_1, \sigma) \propto -nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Maximum Likelihood Estimation

Recall that if $y|x \sim \mathcal{N}(h(x), \sigma^2)$ where $h : \mathbb{R} \to \mathbb{R}$, then the conditional PDF is given by

$$p(y|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-h(x)}{\sigma}\right)^2}.$$

In simple linear regression, recall that $h(x) = \beta_0 + \beta_1 x$. The **(conditional) likelihood function** is given by

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1, \sigma) &\equiv \mathcal{L}(\beta_0, \beta_1, \sigma; \mathcal{D}) \\
&= p(y_1, \ldots, y_n | x_1, \ldots, x_n; \beta_0, \beta_1, \sigma) \\
&= \Pi_{i=1}^{n} p(y_i | x_1, \ldots, x_n; \beta_0, \beta_1, \sigma) \\
&= \Pi_{i=1}^{n} p_{y|x}(y_i | x_i; \beta_0, \beta_1, \sigma) \\
&\propto \sigma^{-n} exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}.
\end{aligned}$$

The **(conditional) log-likelihood** is given by

$$\log \mathcal{L}(\beta_0, \beta_1, \sigma) \propto -nlog(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Maximum Likelihood Estimation

To find the MLE of β_0 and β_1, we **maximize** the conditional log-likelihood:

$$\underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\text{Maximize}} \log \mathcal{L}(\beta_0, \beta_1, \sigma)$$

$$\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{Maximize}\ nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2,$$

$$\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{Maximize}\ -\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{Minimize}\ \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

In other words, **(ordinary) least squares (OLS) is equivalent to MLE** with a linear model and a normally distributed error term.

Maximum Likelihood Estimation

To find the MLE of β_0 and β_1, we **maximize** the conditional log-likelihood:

$$\begin{aligned}
&\underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\text{Maximize}} \log \mathcal{L}(\beta_0,\beta_1,\sigma) \\
&\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\mathit{Maximize}}\ nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2, \\
&\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\mathit{Maximize}}\ -\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\
&\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\mathit{Minimize}}\ \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2
\end{aligned}$$

In other words, **(ordinary) least squares (OLS) is equivalent to MLE** with a linear model and a normally distributed error term.

Maximum Likelihood Estimation

To find the MLE of β_0 and β_1, we **maximize** the conditional log-likelihood:

$$\underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\text{Maximize}} \log \mathcal{L}(\beta_0, \beta_1, \sigma)$$

$$\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\mathit{Maximize}} \; nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2,$$

$$\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\mathit{Maximize}} \; -\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\mathit{Minimize}} \; \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

In other words, **(ordinary) least squares (OLS) is equivalent to MLE** with a linear model and a normally distributed error term.

Maximum Likelihood Estimation

To find the MLE of β_0 and β_1, we **maximize** the conditional log-likelihood:

$$\begin{aligned}
&\underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\text{Maximize}} \log \mathcal{L}(\beta_0,\beta_1,\sigma) \\
&\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\mathit{Maximize}}\ nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2, \\
&\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\mathit{Maximize}}\ -\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2 \\
&\equiv \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\mathit{Minimize}}\ \sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2
\end{aligned}$$

In other words, **(ordinary) least squares (OLS) is equivalent to MLE** with a linear model and a normally distributed error term.

Outline

Simple linear regression

Optimal predictions

Parameter estimation with least squares

Parameter estimation with MLE

Model accuracy and hypothesis testing

Multiple linear regression

Parameter estimation

Interpreting regression coefficients

Qualitative/categorical variables

Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

How to estimate the out-of-sample error in linear regression?

Variable subset selection methods

Shrinkage methods

Assessing the accuracy of the coefficient estimates

The **standard erro**r of an estimator reflects how it varies under repeated sampling. We have

$$\mathsf{SE}(\hat{\beta}_1)^2 = \mathsf{Var}(\hat{\beta}_1) \text{ and } \mathsf{SE}(\hat{\beta}_0)^2 = \mathsf{Var}(\hat{\beta}_0).$$

We can also compute **confidence intervals** (CI). A 95% CI is defined as an interval such that with 95% probability,it will contain the true unknown value of the coefficient. There is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \times \mathsf{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \times \mathsf{SE}(\hat{\beta}_1)\right],$$

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample).

Hypothesis testing

The most common hypothesis test involves testing the **null hypothesis** of

- H_0: There is no relationship between x and y, i.e. $\beta_1 = 0$,

versus the **alternative hypothesis**

- H_A: There is some relationship between x and y, i.e. $\beta_1 \neq 0$,

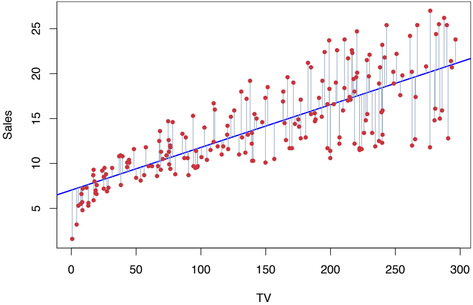
Assuming a linear DGP, i.e. $y = \beta_0 + \beta_1 x + \varepsilon$, if $\beta_1 = 0$, the model reduces to $y = \beta_0 + \varepsilon$ and x is not associated with y.

To test the null hypothesis, we compute a **t-statistics**

$$\frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}.$$

Under the null ($\beta_1 = 0$), we know the t-statistics has a **t-distribution** with $n-2$ degrees of freedom. As a result, we can compute the probability of observing any value equal to $|t|$ or larger (called the **p-value**).

Example



	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Other accuracy measures

The **R-squared** or the fraction of variance explained is defined as

$$\mathrm{R}^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

where TSS is the total sum of squares $\sum_{i=1}^n (y_i - \bar{y})^2$ and RSS is the residual sum of squares $\sum_{i=1}^n (y_i - \hat{y}_i)^2$.

- If $\hat{y}_i = y_i \implies \mathrm{R}^2 = 1$
- If $\hat{y}_i = \bar{y} \implies \mathrm{R}^2 = 0$

The **Residual standard error** (RSE) is defined as

$$RSE = \sqrt{\frac{1}{n-2}\mathrm{RSS}}.$$

Outline

Simple linear regression

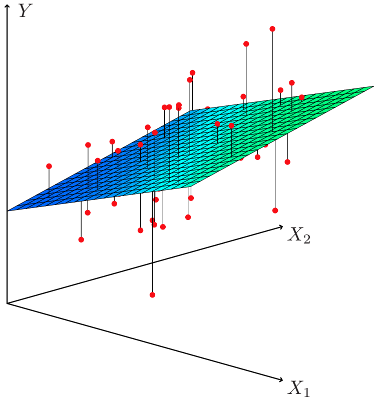
Multiple linear regression

- Parameter estimation
- Interpreting regression coefficients
- Qualitative/categorical variables
- Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

Multiple linear regression

In multiple linear regression, we consider a multivariate input $x \in \mathbb{R}^p$ where $p > 1$. The following figure shows an example with $p = 2$.



Outline

Simple linear regression

Optimal predictions

Parameter estimation with least squares

Parameter estimation with MLE

Model accuracy and hypothesis testing

Multiple linear regression

Parameter estimation

Interpreting regression coefficients

Qualitative/categorical variables

Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

How to estimate the out-of-sample error in linear regression?

Variable subset selection methods

Shrinkage methods

Matrix notation

A dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ where $x_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^T \in \mathbb{R}^{p+1}$ can be represented as

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n,$$

and

$$\boldsymbol{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ 1 & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \in \mathbb{R}^{n\times(p+1)}$$

Parameter estimation - Matrix notation

Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$., the residual sum of squares (RSS) can be written as

$$\text{RSS}(\boldsymbol{\beta}) = \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} = (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta),$$

Assuming $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, we have

$$\hat{\beta} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\text{argmin}} \; (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

with the fitted values given by $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$.

Note: $(\boldsymbol{X}^T\boldsymbol{X})$ is not always invertible, e.g. in high dimensions ($p > n$) or when some input variables are highly correlated.

Parameter estimation - Matrix notation

Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$., the residual sum of squares (RSS) can be written as

$$\text{RSS}(\boldsymbol{\beta}) = \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} = (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta),$$

Assuming $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, we have

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\text{argmin}} \, (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

with the fitted values given by $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$.

Note: $(\boldsymbol{X}^T\boldsymbol{X})$ is not always invertible, e.g. in high dimensions $(p > n)$ or when some input variables are highly correlated.

Outline

Simple linear regression

Optimal predictions

Parameter estimation with least squares

Parameter estimation with MLE

Model accuracy and hypothesis testing

Multiple linear regression

Parameter estimation

Interpreting regression coefficients

Qualitative/categorical variables

Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

How to estimate the out-of-sample error in linear regression?

Variable subset selection methods

Shrinkage methods

Interpreting regression coefficients

- Claims of **causality** should be avoided for observational data.
- The ideal scenario is when the **variables are uncorrelated**. We can interpret the coefficients as follows:
  'a unit change in x_j is associated with a β_j change in y , while all the other variables stay fixed"
- There are issues when the **variables are correlated**:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous: when x_j changes, everything else changes.

Interpreting regression coefficients

- Claims of **causality** should be avoided for observational data.
- The ideal scenario is when the **variables are uncorrelated**. We can interpret the coefficients as follows:
  'a unit change in x_j is associated with a β_j change in y , while all the other variables stay fixed"
- There are issues when the **variables are correlated**:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous: when x_j changes, everything else changes.

Interpreting regression coefficients

- Claims of **causality** should be avoided for observational data.
- The ideal scenario is when the **variables are uncorrelated**. We can interpret the coefficients as follows:
  ‘a unit change in x_j is associated with a β_j change in y , while all the other variables stay fixed”
- There are issues when the **variables are correlated**:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous: when x_j changes, everything else changes.

Example

$$\textit{sales} = \beta_0 + \beta_1 \times \textit{TV} + \beta_2 \times \textit{radio} + \beta_3 \times \textit{newspaper}.$$

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Outline

Simple linear regression

- Optimal predictions
- Parameter estimation with least squares
- Parameter estimation with MLE
- Model accuracy and hypothesis testing

Multiple linear regression

- Parameter estimation
- Interpreting regression coefficients
- Qualitative/categorical variables
- Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

- How to estimate the out-of-sample error in linear regression?
- Variable subset selection methods
- Shrinkage methods

Discrete/qualitative variables

- Some predictors are not quantitative but are **qualitative**, taking a **discrete** set of values (levels), also called categorical predictors or factor variables.
- For these, we need to create **dummy variables**. There will always be one fewer dummy variable than the number of levels. The level with no dummy variable is known as the **baseline**.
- Let us consider two examples where we regress the credit card balance (y) on either gender (male or female) or ethnicity (Caucasian, African American (AA) or Asian)

Discrete/qualitative variables

- Some predictors are not quantitative but are **qualitative**, taking a **discrete** set of values (levels), also called categorical predictors or factor variables.
- For these, we need to create **dummy variables**. There will always be one fewer dummy variable than the number of levels. The level with no dummy variable is known as the **baseline**.
- Let us consider two examples where we regress the credit card balance (y) on either gender (male or female) or ethnicity (Caucasian, African American (AA) or Asian)

Discrete/qualitative variables

- Some predictors are not quantitative but are **qualitative**, taking a **discrete** set of values (levels), also called categorical predictors or factor variables.
- For these, we need to create **dummy variables**. There will always be one fewer dummy variable than the number of levels. The level with no dummy variable is known as the **baseline**.
- Let us consider two examples where we regress the credit card balance (y) on either gender (male or female) or ethnicity (Caucasian, African American (AA) or Asian)

Example I

Let us regress the credit card balance (y) on gender (male or female). We create a new **dummy** variable

$$x = \begin{cases} 1, & \text{if the person is female} \\ 0, & \text{if the person is male } (\textit{baseline}) \end{cases}$$

Our model is

$$h(x) = \beta_0 + \beta_1 x = \begin{cases} \beta_0 + \beta_1, & \text{if the person is female} \\ \beta_0, & \text{if the person is male} \end{cases}$$

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

Example II

Let us regress the credit card balance (y) on ethnicity (Caucasian, African American (AA) or Asian). When the variable takes three values, we need two dummy variables. Then, our model is

$$h(x) = \beta_0 + \beta_1 x + \beta_2 x = \begin{cases} \beta_0 + \beta_1, & \text{if the person is Asian} \\ \beta_0 + \beta_2, & \text{if the person is Caucasian} \\ \beta_0, & \text{if the person is AA } (\textit{baseline}) \end{cases}$$

	Coefficient	Std. Error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
ethnicity[Asian]	-18.69	65.02	-0.287	0.7740
ethnicity[Caucasian]	-12.50	56.68	-0.221	0.8260

Outline

Simple linear regression

Optimal predictions
Parameter estimation with least squares
Parameter estimation with MLE
Model accuracy and hypothesis testing

Multiple linear regression

Parameter estimation
Interpreting regression coefficients
Qualitative/categorical variables
Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

How to estimate the out-of-sample error in linear regression?
Variable subset selection methods
Shrinkage methods

Extensions of the linear model

- The linear model is **linear in the variables**. We can extend the linear model by applying **transformations** to the variables. While doing so, the model remains linear in the variables.
- We will consider two extensions: **interactions** and **non-linearity**

Interaction effect

$$\textit{sales} = \beta_0 + \beta_1 \times \textit{TV} + \beta_2 \times \textit{radio}$$

- We assumed that the effect on sales of increasing one advertising medium is **independent** of the amount spent on the other media
- For example, the average effect on sales of a one-unit increase in TV is always β_1, regardless of the amount spent on radio.
- But suppose that spending money on radio advertising **actually increases the effectiveness** of TV advertising, so that the slope term for TV should increase as radio increases.
- In marketing, this is known as a synergy effect, and in statistics it is referred to as an i**nteraction effect**.

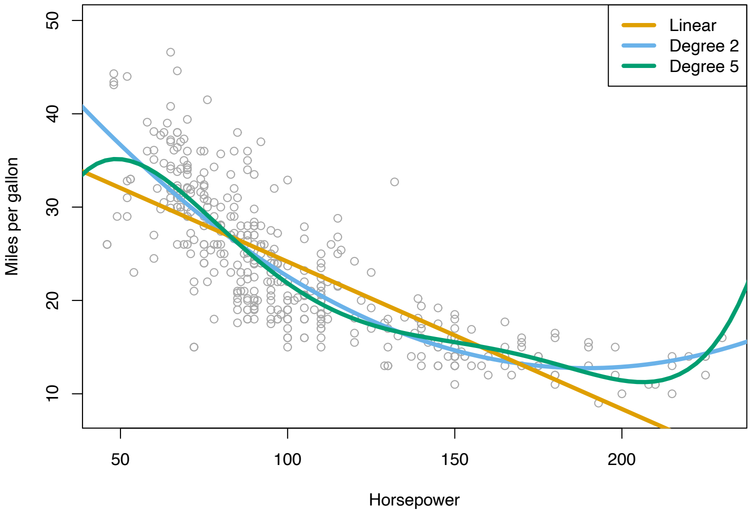
Interaction effect

We can model interactions as follows:

$$\begin{aligned} \textit{sales} &= \beta_0 + \beta_1 \times \textit{TV} + \beta_2 \times \textit{radio} + \beta_3 \times (\textit{radio} \times \textit{TV}) \qquad (3) \\ &= \beta_0 + (\beta_1 + \beta_3 \times \textit{radio}) \times \textit{TV} + \beta_2 \times \textit{radio} \qquad (4) \end{aligned}$$

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Nonlinear effect



Nonlinear effect

$$\mathit{mpg} = \beta_0 + \beta_1 \times \mathit{horsepower} + \beta_2 \times \mathit{horsepower}^2$$

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	56.9001	1.8004	31.6	< 0.0001
`horsepower`	-0.4662	0.0311	-15.0	< 0.0001
`horsepower`^{2}	0.0012	0.0001	10.1	< 0.0001

Topics not covered

- Outliers
- Non-constant variance of error terms
- High leverage points
- Collinearity

Outline

Simple linear regression

Multiple linear regression

Linear Model Selection and Regularization

How to estimate the out-of-sample error in linear regression?

Variable subset selection methods

Shrinkage methods

Linear model selection

- With linear models, we need to select the **best subset of input variables**.
- By removing irrelevant variables, we can obtain a model that provides **better predictions** (less variance) and is more **easily interpretable**.
- If there are a limited number of predictors, we can consider all possible models. Otherwise we need a **search strategy** to explore some potential models.
- Let us first discuss different methods to estimate the out-of-sample error of a given linear model.

Outline

Simple linear regression

Optimal predictions
Parameter estimation with least squares
Parameter estimation with MLE
Model accuracy and hypothesis testing

Multiple linear regression

Parameter estimation
Interpreting regression coefficients
Qualitative/categorical variables
Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

How to estimate the out-of-sample error in linear regression?
Variable subset selection methods
Shrinkage methods

How to estimate the out-of-sample error?

$$E_{\mathrm{out}}(h) = E_{\mathrm{in}}(h) + \underbrace{[E_{\mathrm{out}}(h) - E_{\mathrm{in}}(h)]}_{\textbf{overfit penalty}}, \quad h \in \mathcal{H}.$$

1. **Directly estimate it** using resampling methods.
2. **Estimate the overfit penalty/optimism** and **add it to the in-sample (training) error**.

Leave-one-out cross-validation with linear models

Let $\hat{y}_{[i]}$ be the predicted value obtained when the model is estimated with the ith observation deleted. If $e_{[i]} = y_i - \hat{y}_{[i]}$, then the leave-one-out cross-validation error is given by

$$E_{\text{loo}} = \frac{1}{n}\sum_{i=1}^{n} e_{[i]}^2,$$

It turns out that for linear models, we **do not** actually have to estimate the model n times, once for each omitted case.

Recall that $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. The fitted values are $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{H}\boldsymbol{y}$ with $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. If the diagonal values of $\boldsymbol{H}$ are denoted by $h_1, \dots, h_n$, then we have

$$E_{\text{loo}} = \frac{1}{n}\sum_{i=1}^{n}[e_i/(1-h_i)]^2,$$

where $e_i = y_i - \hat{y}_i$.

Leave-one-out cross-validation with linear models

Let $\hat{y}_{[i]}$ be the predicted value obtained when the model is estimated with the ith observation deleted. If $e_{[i]} = y_i - \hat{y}_{[i]}$, then the leave-one-out cross-validation error is given by

$$E_{\text{loo}} = \frac{1}{n}\sum_{i=1}^{n} e_{[i]}^2,$$

It turns out that for linear models, we **do not** actually have to estimate the model n times, once for each omitted case.

Recall that $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. The fitted values are $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{H}\boldsymbol{y}$ with $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. If the diagonal values of $\boldsymbol{H}$ are denoted by $h_1, \ldots, h_n$, then we have

$$E_{\text{loo}} = \frac{1}{n}\sum_{i=1}^{n}[e_i/(1-h_i)]^2,$$

where $e_i = y_i - \hat{y}_i$.

Training error adjustment

- One advantage of resampling methods is the fact that they can be used in a wider range of model selection tasks, even in cases where it is **hard to pinpoint the “model size”**.
- With linear models, there are various methods based on ”training error adjustment” since it is **easier** to estimate ”model size”.

Expected in-sample vs out-of-sample errors

Let us compare the **expected** in-sample and out-of-sample **MSE** in a specific scenario. We assume that the training data is given by

$$\{(x_i, y_i)\}_{i=1}^{n} \text{ with } y_i = f(x_i) + \varepsilon_i,$$

and the test data is given by

$$\{(x_i, y_i^{'})\}_{i=1}^{n} \text{ with } y_i^{'} = f(x_i) + \varepsilon_i^{'},$$

where x_i are fixed (not random) and ε_i and $\varepsilon_i^{'}$ are **independent** but **identically** distributed random noise variables.

In other words, the training and test data share the **same** input variables x_i but have **different** random noise terms. This scenario is a particular case (simpler to analyze) of the more general scenario where the x_i in the training and test data can be different.

Expected in-sample vs out-of-sample errors

Let $\hat{y}_i = g(x_i)$ where g is computed using the training data $\{(x_i, y_i)\}_{i=1}^n$. We can show that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i' - \hat{y}_i]^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2\right] + \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(y_i, \hat{y}_i)$$

The last term in the RHS. of the previous expression is called the **optimism**, which is the amount by which the training error systematically under-estimates the expected test error.

If we assume the data generating process is **linear** and if we use the **least square estimator**, we can show that

$$\sum_{i=1}^{n}\mathrm{Cov}(y_i, \hat{y}_i) = \sigma^2(p+1),$$

The problem with Residual Sum of Squares and R^2

Recall that the **Residual Sum of Squares** (or RSS) is given by

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Minimizing RSS will always choose the model with the most predictors.

Recall that the R^2 **statistic** is given by

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

The R^2 gives the proportion of variance explained, and is independent of the scale of y. However . . .

- R^2 does not allow for "degrees of freedom".
- Adding *any* variable tends to increase the value of R^2, even if that variable is irrelevant.

The problem with Residual Sum of Squares and R^2

Recall that the **Residual Sum of Squares** (or RSS) is given by

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Minimizing RSS will always choose the model with the most predictors.
Recall that the R^2 **statistic** is given by

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

The R^2 gives the proportion of variance explained, and is independent of the scale of y. However . . .

- R^2 does not allow for "degrees of freedom".
- Adding *any* variable tends to increase the value of R^2, even if that variable is irrelevant.

Estimated residual variance and adjusted R^2

Insead of minimizing RSS, we can minimize the **estimated residual variance**, given by

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n - \boldsymbol{p} - 1},$$

where p is the number of predictors.

Also, instead of R^2, we can use the **adjusted** R^2, defined by

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n - \boldsymbol{p} - 1} = 1 - \frac{\text{RSS}/(n - \boldsymbol{p} - 1)}{\text{TSS}/(n-1)},$$

which pays a price for the inclusion of unnecessary variables.

Maximizing $\bar{R}^2$ is equivalent to minimizing $\hat{\sigma}^2$.

Estimated residual variance and adjusted R^2

Minimizing $\hat{\sigma}^2$, what does that translate to? We have

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-p-1} = \mathrm{MSE}\frac{n}{n-p-1} = \mathrm{MSE}\frac{1}{1-(p+1)/n}.$$

Using the binomial theorem which gives $(1-x)^{-1} = 1 + x + x^2 + \dots$, and truncating the series at first order[1], we obtain

$$\hat{\sigma}^2 \approx \mathrm{MSE}\left(1 + \frac{p+1}{n}\right) = \mathrm{MSE} + \mathrm{MSE}\frac{p+1}{n}.$$

Even for the right model (where MSE is a consistent estimator of σ^2), the penalty is half as big as what it should be, i.e.

$$\mathrm{MSE} + 2 \times \sigma^2 \frac{(p+1)}{n}.$$

$\implies$ $\bar{R}^2$ is better than R^2 but it is still not going to work very well.

[1] For a fixed p, the approximation becomes exact as $n \to \infty$.

Estimated residual variance and adjusted R^2

Minimizing $\hat{\sigma}^2$, what does that translate to? We have

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p-1} = \text{MSE}\frac{n}{n-p-1} = \text{MSE}\frac{1}{1-(p+1)/n}.$$

Using the binomial theorem which gives $(1-x)^{-1} = 1 + x + x^2 + \dots$, and truncating the series at first order[1], we obtain

$$\hat{\sigma}^2 \approx \text{MSE}\left(1 + \frac{p+1}{n}\right) = \text{MSE} + \text{MSE}\frac{p+1}{n}.$$

Even for the right model (where MSE is a consistent estimator of σ^2), the penalty is half as big as what it should be, i.e.

$$\text{MSE} + 2 \times \sigma^2 \frac{(p+1)}{n}.$$

$\Longrightarrow$ $\bar{R}^2$ is better than R^2 but it is still not going to work very well.

[1] For a fixed p, the approximation becomes exact as $n \to \infty$.

Mallow's C_p

The Mallows C_p statistic is given by

$$C_p = \frac{1}{n}(RSS + 2(p+1)\hat{\sigma}^2),$$

where p is the number of predictors in the model.

It essentially substitutes an estimator of σ^2 in the expression of the **optimism** for linear models. C_p penalizes more heavily than $\bar{R}^2$.

Akaike's Information Criterion

$$\text{AIC} = -2\log(\mathcal{L}) + 2(p+1)$$

where $\mathcal{L}$ is the likelihood and p is the number of predictors.

- AIC is defined for a large class of models fit by **maximum likelihood estimation**. It is also called a **penalized likelihood** approach.
- In the case of the **linear model with Gaussian errors**, maximum likelihood and least squares are the same thing, and C_p and AIC are equivalent.
- AIC is **asymptotically** equivalent to leave-one-out cross-validation.
- *Minimizing* the AIC gives the best model for **prediction** (not inference).

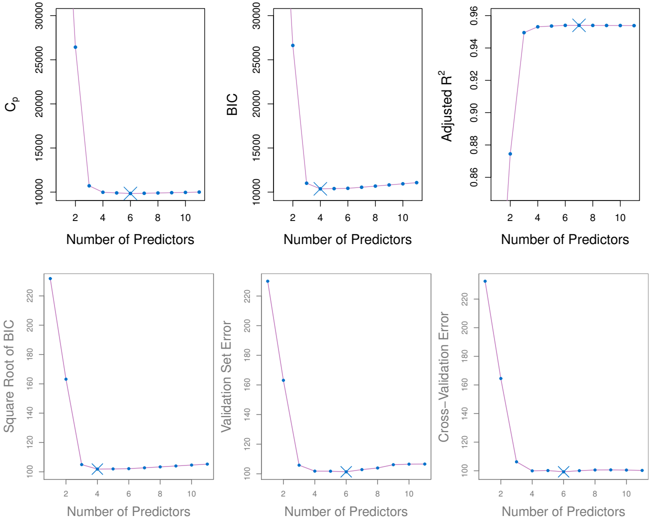
Schwartz Bayesian Information Criterion

$$\mathrm{BIC} = -2\log(\mathcal{L}) + (p+1)\log(n)$$

where $\mathcal{L}$ is the likelihood and p is the number of predictors.

- BIC penalizes more **heavily** than AIC
- Since $log(n) > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of **smaller models** than C_p/AIC.
- Also called SBIC and SC.
- BIC is **asymptotically** equivalent to leave-v-out cross-validation when $v = n[1 - 1/(log(n) - 1)]$.

Example


Cp
10000
15000
20000
25000
30000
2
4
6
8
10
Number of Predictors
BIC
10000
15000
20000
25000
30000
2
4
6
8
10
Number of Predictors
Adjusted R2
0.86
0.88
0.90
0.92
0.94
0.96
2
4
6
8
10
Number of Predictors
Square Root of BIC
100
120
140
160
180
200
220
2
4
6
8
10
Number of Predictors
Validation Set Error
100
120
140
160
180
200
220
2
4
6
8
10
Number of Predictors
Cross−Validation Error
100
120
140
160
180
200
220
2
4
6
8
10
Number of Predictors

Outline

Simple linear regression

Optimal predictions

Parameter estimation with least squares

Parameter estimation with MLE

Model accuracy and hypothesis testing

Multiple linear regression

Parameter estimation

Interpreting regression coefficients

Qualitative/categorical variables

Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

How to estimate the out-of-sample error in linear regression?

Variable subset selection methods

Shrinkage methods

Variable subset selection

- When performing (linear) model selection, we often need to select a **subset of the input variables**.
- Removing (irrelevant) variables can yield better **prediction accuracy** and **model interpretability**.
- If there are a limited number of predictors, we can study all possible models. Otherwise we need a **search strategy** to explore some potential models.
- Although we will present selection strategies for linear regression models, the same ideas apply to **other types of models**.
- in the following, we will present **best subset** and **stepwise** model selection procedures.

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.

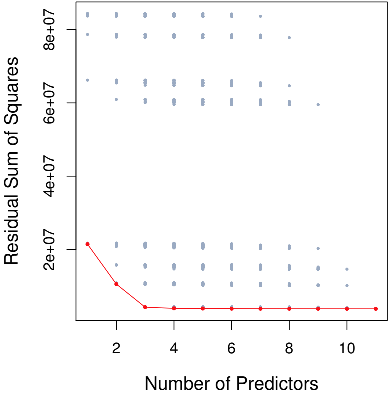
2. For $k = 1, 2, \ldots p$:

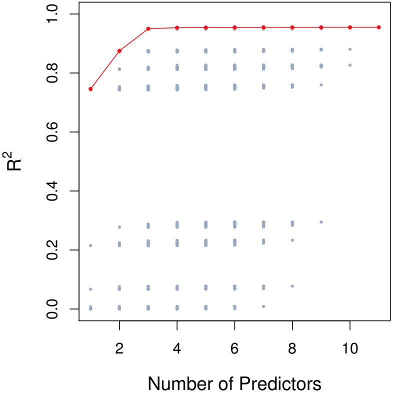
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

   (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.

Best subset selection





An example with $p = 10$ variables including a categorical variable with three categories.

Best subset selection

- Best subset selection may suffer from both **computational** and **overfitting** problems with a large number of variables p.
- Best subset selection will fit and evaluate 2^p models.
- A larger search space increase the chance of finding models that look good on the training data, while being inaccurate on new data. In other words, a large search space can lead to **overfitting** and **high variance** of the coefficient estimates.
- **Stepwise** methods are attractive alternatives to best subset selection since they explore a more **restricted set** of models.

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

- Forward stepwise selection begins with a model containing no predictors. At each step, the variable that gives the greatest **additional** improvement to the fit is added to the previously selected variables.
- It is **less** computationally demanding that best subset selection. It searches through $1 + p(p+1)/2$ models
- It is **not guaranteed** to find the best possible model out of all 2^p models containing subsets of the p variables.

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

- Backward stepwise selection begins with the **full** least squares model containing all p variables, and then iteratively removes the **least useful** variable, one-at-a-time.
- Like forward stepwise selection, it is **less** computationally demanding that best subset selection, and it is **not guaranteed** to yield the best model containing a subset of the p predictors.
- It requires that the number of samples n is **larger** than the number of variables p (so that the full model can be fit). Forward stepwise does not have this limitation and can be used when $p >> n$.

Outline

Simple linear regression

Optimal predictions

Parameter estimation with least squares

Parameter estimation with MLE

Model accuracy and hypothesis testing

Multiple linear regression

Parameter estimation

Interpreting regression coefficients

Qualitative/categorical variables

Extensions of the linear model: interactions and non-linear effects

Linear Model Selection and Regularization

How to estimate the out-of-sample error in linear regression?

Variable subset selection methods

Shrinkage methods

Best subset selection

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0) \leq s,$$

where $\mathbb{I}(\cdot)$ is the indicator function and $s \geq 0$ is a **hyperparameter**.

- This problem is equivalent to best subset selection. We search for a set of coefficient estimates such that RSS is as small as possible, subject to the constraint that no more than s coefficients can be nonzero
- Recall that it is computationally infeasible when p is large, since it requires considering all $\binom{p}{s}$ models containing s predictors
- We can use stepwise procedures but the search space becomes restricted.

Shrinkage methods

- Subset Selection: we identify the **best subse**t of the p predictors. We then fit a model using least squares on the reduced set of variables.
- Shrinkage: We fit a model involving all p predictors, but the estimated coefficients are **shrunken towards zero** relative to the least squares estimates. This shrinkage (also known as **regularization**) has the effect of **reducing variance** and can also perform **variable selection**.
- In the following, we will present two shrinkage methods: **Ridge regression** and **the LASSO**[2].

[2] least absolute shrinkage and selection operator

Ridge regression

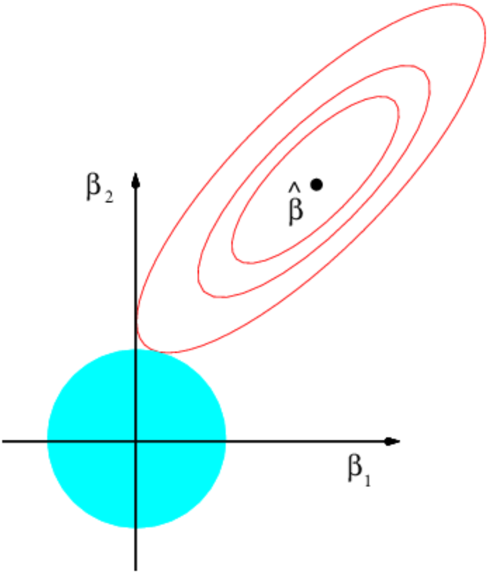
$$\underset{\boldsymbol{\beta}}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq s,$$

where $s \geq 0$ is a **hyperparameter**.

$$\begin{aligned} s = 0 \quad & \rightarrow \quad \hat{\boldsymbol{\beta}}^{\mathrm{R}} = (0, \ldots, 0) \\ s \rightarrow \infty \quad & \rightarrow \quad \hat{\boldsymbol{\beta}}^{\mathrm{R}} \rightarrow \hat{\boldsymbol{\beta}}^{\mathrm{ls}} \text{ (least squares)} \\ s \in (0, \infty) \quad & \rightarrow \quad \text{bias-variance tradeoff} \end{aligned}$$

Ridge regression: geometry



Ridge regression: Lagrangian form

$$\begin{aligned} & \underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 && (5) \\ \equiv \; & \underset{\boldsymbol{\beta}}{\text{minimize}} \; \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda \; \|\boldsymbol{\beta}\|_2^2, && (6) \end{aligned}$$

where $\lambda \geq 0$ is a **hyperparameter**.

$$\begin{aligned} \lambda = 0 \quad & \rightarrow \quad \hat{\boldsymbol{\beta}}^{\mathrm{R}} = \hat{\boldsymbol{\beta}}^{\mathrm{ls}} \\ \lambda \rightarrow \infty \quad & \rightarrow \quad \hat{\boldsymbol{\beta}}^{\mathrm{R}} \rightarrow (0, \ldots, 0) \\ \lambda \in (0, \infty) \quad & \rightarrow \quad \text{bias-variance tradeoff} \end{aligned}$$

Note that the ridge objective function can be written as a standard least squares objective after *data augmentation*.

A Simple special case

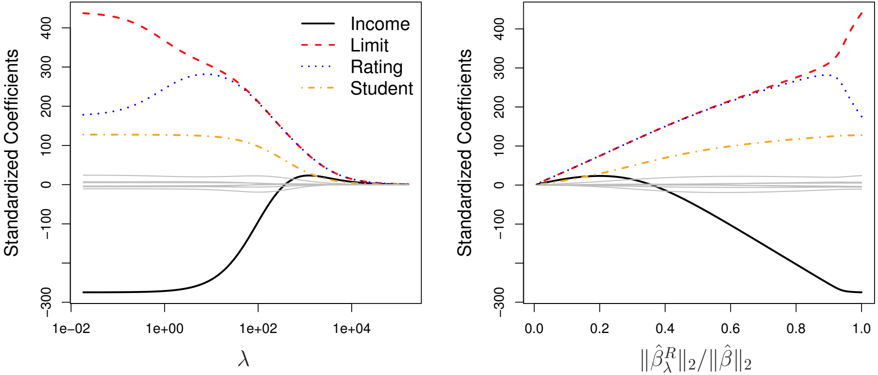
Let us consider a special case where $n = p$ and $x_{ij} = 1$ if $i = j$ and $x_{ij} = 0$ otherwise. In other words, $\boldsymbol{X}$ is an identity matrix.

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{j=1}^{p} (y_j - \beta_j)^2 \quad \rightarrow \quad \hat{\beta}_j^{\text{ls}} = y_j$$

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad \rightarrow \quad \hat{\beta}_j^{R} = \frac{y_j}{(1+\lambda)} = \frac{\hat{\beta}_j^{ls}}{(1+\lambda)}$$

This illustrates the essential feature of ridge regression: shrinkage. Furthermore, we can see that ridge regression **introduces bias but reduces the variance**.

Ridge regression example

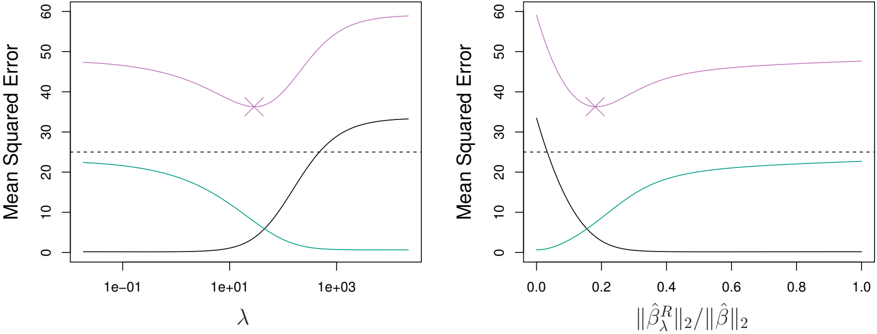


While the ridge coefficient estimates tend to **decrease in aggregate** as λ increases, individual coefficients, such as rating and income, may **occasionally increase** as λ increases.

A note on scaling

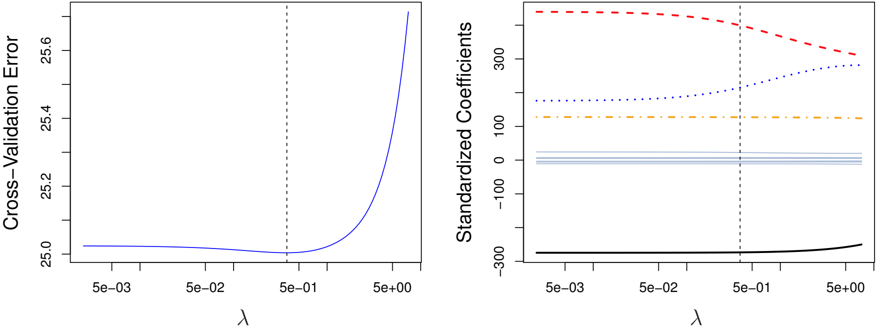
- Standard least squares coefficient estimates are **scale equivariant**
  - multiplying X_j by a constant c simply leads to a scaling of the least squares coefficient estimates by a factor of $1/c$
  - regardless of how the jth predictor is scaled, $X_j\hat{\beta}_j$ will remain the same.
- The ridge regression coefficient estimates **can change substantially** when multiplying a given predictor by a constant
  - This is due to the sum of squared coefficients term in the ridge regression formulation
  - If we use thousands of dollars instead of dollars, it will **not** simply cause the ridge estimate to change by a factor of $1,000$

Ridge Regression vs Least Squares

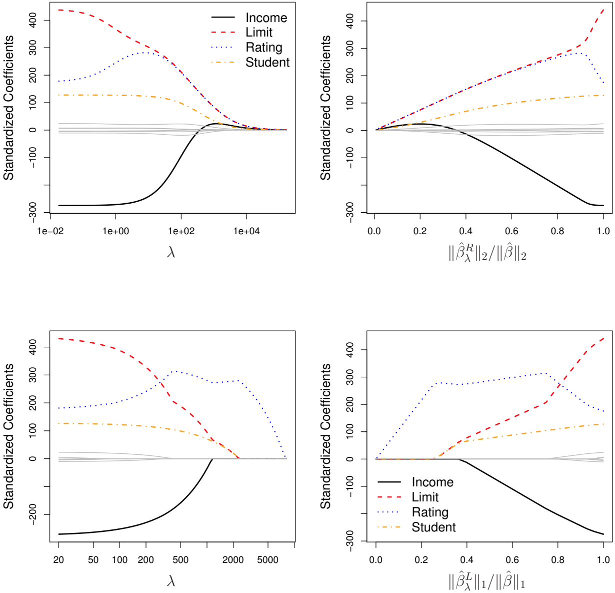


Squared bias (black), variance (green), and test mean squared error (purple)

Selecting the Tuning Parameter



Another shrinkage method



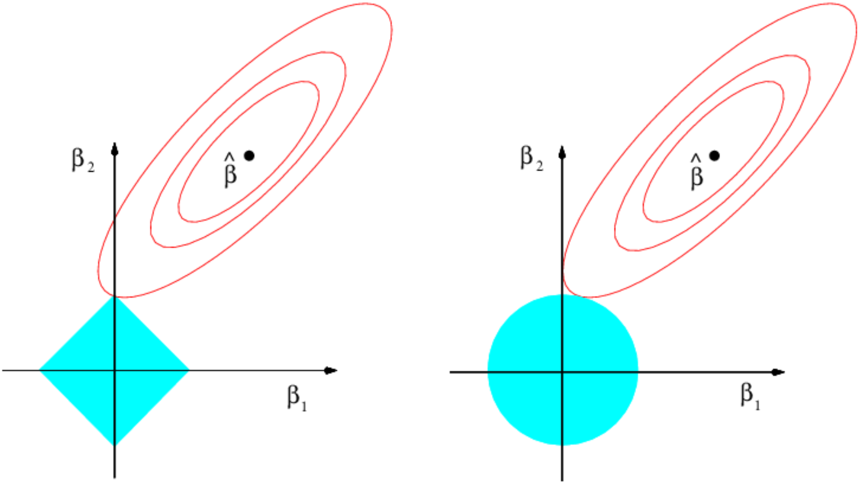
LASSO regression

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$
$$\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq s,$$

where $s \geq 0$ is a **hyperparameter**.

$$\begin{aligned} s &= 0 && \rightarrow && \hat{\boldsymbol{\beta}}^{\text{L}} = (0, \ldots, 0) \\ s &\rightarrow \infty && \rightarrow && \hat{\boldsymbol{\beta}}^{\text{L}} \rightarrow \hat{\boldsymbol{\beta}}^{\text{ls}} \text{ (least squares)} \\ s &\in (0, \infty) && \rightarrow && \text{bias-variance tradeoff} \end{aligned}$$

LASSO vs Ridge geometry



LASSO: Lagrangian form

$$\begin{aligned} & \underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| && (7) \\ \equiv\ & \underset{\boldsymbol{\beta}}{\text{minimize}}\ \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2 + \lambda\ \|\boldsymbol{\beta}\|_1, && (8) \end{aligned}$$

where $\lambda \geq 0$ is a **hyperparameter**.

$$\begin{aligned} \lambda = 0 \quad & \rightarrow \quad \hat{\boldsymbol{\beta}}^{\mathsf{L}} = \hat{\boldsymbol{\beta}}^{\mathsf{ls}} \\ \lambda \rightarrow \infty \quad & \rightarrow \quad \hat{\boldsymbol{\beta}}^{\mathsf{L}} \rightarrow (0, \ldots, 0) \\ \lambda \in (0, \infty) \quad & \rightarrow \quad \text{bias-variance tradeoff} \end{aligned}$$

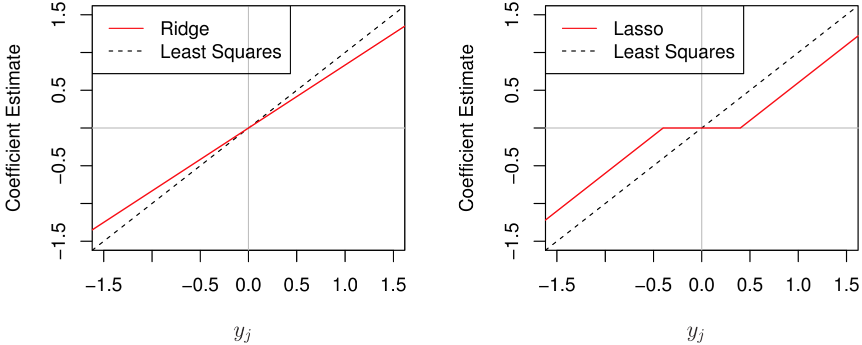
A Simple special case with LASSO

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$\hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2} & \text{if } y_j > \frac{\lambda}{2}; \\ y_j + \frac{\lambda}{2} & \text{if } y_j < -\frac{\lambda}{2}; \\ 0 & \text{if } |y_j| \leq \frac{\lambda}{2}. \end{cases}$$

The lasso shrinks each least squares coefficient towards zero by a **constant amount**, $\lambda/2$. The least squares coefficients that are less than $\lambda/2$ in absolute value are **shrunken entirely to zero**.

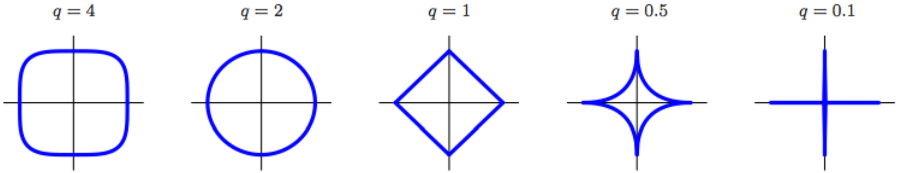
A Simple special case



Note: For Least Squares, we have $\beta_j = y_j$

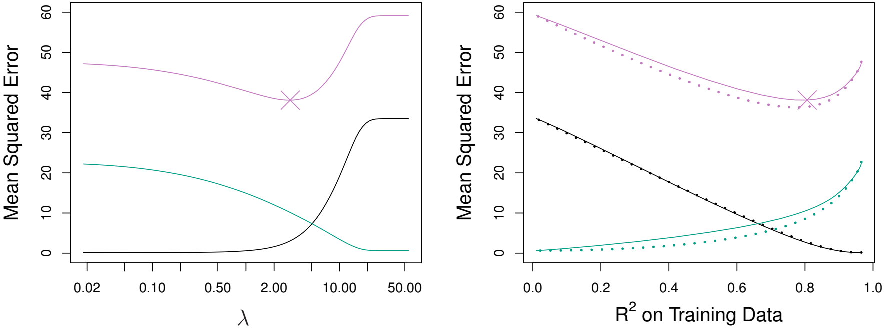
LASSO and Sparsity

- A vector $\beta \in \mathbb{R}^p$ is k-sparse if it has **at most k nonzero entries**.
- q-norm regularization with $q > 1$ does not provide sparse coefficient estimate, e.g. ridge regression
- For $q < 1$, the solutions are sparse but the problem is **not convex** and this makes the optimisation very challenging computationally.
- The value $q = 1$ (LASSO) is the smallest value that yields a **convex problem**.



Lasso vs ridge regression

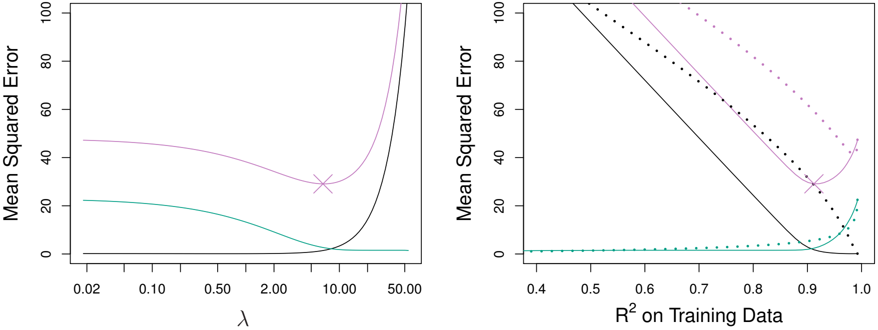
A simulated data set containing p = 45 predictors and n = 50 observations where **all 45 predictors are related to the response**.



Left: Lasso. **Right**: Lasso (solid) and ridge (dashed).

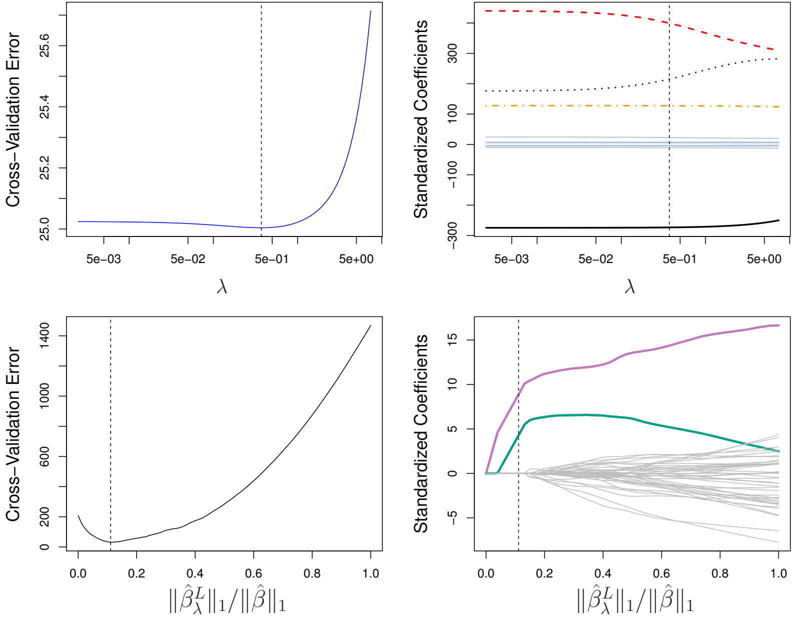
Lasso vs ridge regression

Now the response is a function of **only 2 out of 45 predictors**.



Left: Lasso. **Right**: Lasso (solid) and ridge (dashed).

Selecting the Tuning Parameter


Cross−Validation Error
25.0
25.2
25.4
25.6
5e−03
5e−02
5e−01
5e+00
λ
Standardized Coefficients
−300
−100
0
100
300
5e−03
5e−02
5e−01
5e+00
λ
Cross−Validation Error
0
200
600
1000
1400
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$
Standardized Coefficients
−5
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$