MATH1312: Lecture Note on Probability Theory and Mathematical Statistics

Shixiao W. Jiang

Institute of Mathematical Sciences, ShanghaiTech University, Shanghai 201210, China

jiangshx@shanghaitech.edu.cn

2024 年 11 月 29 日

Contents

Chapter 1

Regression

1.1 General Setup

See references in Cucker and Smale 2001, Bias_Var_Ridge, Learning Theory from First Principles by Francis Bach.

Since we want to study learning from random sampling, the primary object in our development is a probability measure ρ governing the sampling and which is not known in advance (however, the goal is not to reveal ρ).

Let X be a compact domain or a manifold in Euclidean space and $Y = \mathbb{R}^k$. For convenience we will take $k = 1$ for the time being. Let ρ be a Borel probability measure on $Z = X \times Y$ whose regularity properties will be assumed as needed. In the following we try to utilize concepts formed naturally and solely from X, Y and ρ.

A main concept is the **error (or least squares error)** of an arbitrary well-defined functoin f defined by

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho(x, y), \quad \text{for } f : X \to Y. \tag{1.1}$$

For each input $x \in X$ and output $y \in Y$, $(f(x) - y)^2$ is the error suffered from the use of f as a model for the process producing y from x. By integrating over $X \times Y$ (w.r.t. ρ, of course) we average out the error over all pairs (x, y). Hence the word "error" for $\mathcal{E}(f)$.

The problem is posed: What is the f which minimizes the error $\mathcal{E}(f)$?

The error $\mathcal{E}(f)$ naturally decomposes as a sum. Let us see how. For every $x \in X$, let $\rho(y|x)$ be the conditional (w.r.t. x) probability measure on Y and ρ_X be the marginal probability measure on X, i.e. the measure on X defined by $\rho_X(S) = \rho(\pi^{-1}(S))$ where $\pi : X \times Y \to X$ is the projection. Notice that ρ, $\rho(y|x)$ and ρ_X are related as follows. For every integrable function $\varphi : X \times Y \to \mathbb{R}$ a version of Fubini's Theorem states that

$$\int_{X \times Y} \varphi(x, y) d\rho = \int_X \left(\int_Y \varphi(x, y) d\rho(y|x) \right) d\rho_X.$$

This "breaking" of ρ into the measures $\rho(y|x)$ and ρ_X corresponds to looking at Z as a product of an input domain X and an output set Y. In what follows, unless otherwise specified, integrals are to be understood over ρ, $\rho(y|x)$ or ρ_X.

Regression is a method for studying the relationship between a **response variable** Y and a **covariate** X. The covariate is also called a **predictor variable** or a **feature**. One way to summarize the relationship between X and Y is through the **regression function** $f_* : X \to Y$,

$$f_*(x) = E(Y|X = x) = \int_Y y d\rho(y|x).$$

For each $x \in X$, $f_*(x)$ is the average of the y coordinate of $\{x\} \times Y$ (in topological terms, the average of y on the fiber of x). Regularity hypotheses on ρ will induce regularity properties on f_*. We will assume throughout this paper that f_* is bounded. Note that while ρ and f_* are mainly "unknown", ρ_X is known in some situations and can even be the Lebesgue measure on X inherited from Euclidean space. Our goal is to estimate the regression function f_* from the data of the form

$$(Y_1, X_1), \ldots, (Y_n, X_n) \sim F_{X,Y}.$$

Definition 1.1.1 *(Model Assumption for General Setup). The model requires assumptions about how the data are generated. We assume that*

• *there is a "true" function f_* such that the relationship between input and output is for all $i \in \{1, \ldots, n\}$,*

$$y_i = f_*(x) + \epsilon_i. \tag{1.2}$$

*The "true" function f_* can be given as a parametric form such as $f_*(x) = x^\top \theta_*$ (linear regression), $f_*(x) = \varphi(x)^\top \theta_*$ (feature regression), etc. This type of regression is referred to as a **parametric regression**. The function can also be given in a specific form in some function space such as in Sobolev space. We then need to choose a parameterized family of prediction functions $f_\theta : X \to Y$ for $\theta \in \Theta$ in some **high dimensional hypothesis space**. This type of regression is referred to as a **nonparametric regression**. Note that in most cases, the predictor f_* does not belong to the class of functions $\{f_\theta, \theta \in \Theta\}$, that is, the model is said misspecified. These terminologies are not rigorous.*

• *for all $i \in \{1, \ldots, n\}$, ϵ_i are independent such that*

$$\begin{aligned} E(\epsilon_i) &= E(\epsilon_i|x_i) = 0, \\ Var(\epsilon_i) &= Var(\epsilon_i|x_i) = \sigma^2. \end{aligned}$$

Proposition 1.1.2 *For every $f : X \to Y$,*

$$\mathcal{E}(f) = \int_X (f(x) - f_*(x))^2 d\rho_X + \underbrace{\int_Z (f_*(x) - y)^2 d\rho(x, y)}_{\sigma^2}. \tag{1.3}$$

The proof is easily followed by

$$\begin{aligned} \mathcal{E}(f) &= \int_Z (f(x) - y)^2 d\rho(x, y) = \int_Z (f(x) - f_*(x) + f_*(x) - y)^2 d\rho(x, y) \\ &= \int_Z (f(x) - f_*(x))^2 d\rho(x, y) + \int_Z (f_*(x) - y)^2 d\rho(x, y) + \int_Z 2(f(x) - f_*(x))(f_*(x) - y) d\rho(x, y) \\ &= \int_X (f(x) - f_*(x))^2 d\rho_X + \int_Z (f_*(x) - y)^2 d\rho(x, y). \end{aligned}$$

The first term in the right-hand side of Proposition 1.1.2 provides an average (over X) of the error suffered from the use of f as a model for f_*. In addition, since σ^2 is independent of f, Proposition 1.1.2 implies that f_* has the smallest possible error among all functions $f : X \to Y$. Thus σ^2 represents a lower bound on the error $\mathcal{E}(f)$, and it is due solely to our primary object, the measure ρ. Thus, Proposition 1.1.2 supports: **The goal is to "learn" (i.e. to find a good approximation of) f_* from random samples on Z.**

1.2 Simple Linear Regression

In this lecture note, *we only consider the parametric regresssion.* The simplest version of regression is when X_i is simple (one-dimensional) and $f_*(x)$ is assumed to be linear:

$$f_*(x) = \beta_0 + \beta_1 x.$$

This model is called the the **simple linear regression model**. We will make the further simplifying assumption that $Var(\epsilon_i|X = x) = \sigma^2$ does not depend on x. We can thus write the linear regression model as follows.

Definition 1.2.1 *The Simple Linear Regression Model*

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where $E(\epsilon_i|X_i) = 0$ and $Var(\epsilon_i|X_i) = \sigma^2$. The variables β_0 and β_1 are called regression coefficients. In a fixed designed setting, Y is an observable random variable, X is observable fixed non-random variable, and ϵ is unobservable random variables.

Remark 1.2.2 *Warning! Pay attention to the model assumption and model derivation. In the model, whether the distribution of the noise term is specified or only the mean and the variance of the noise term is specified.*

The unknown parameters in the model are the intercept β_0 and the slope β_1 and the variance σ^2. Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ denote estimates of β_0 and β_1. The **fitted line** (or the **hypothesis space**) is

$$\widehat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

The **predicted values** or **fitted values** are $\widehat{Y}_i = \widehat{f}(X_i)$ and the **residuals** are defined to be

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right).$$

The **residual sums of squares** or RSS, which measures how well the line fits the data, is defined by $\text{RSS} = \sum_{i=1}^n \widehat{\epsilon}_i^2$.

Definition 1.2.3 *The **least squares estimates** are the values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize* $\text{RSS} = \sum_{i=1}^n \widehat{\epsilon}_i^2$. *That is*

$$\begin{aligned}(\widehat{\beta}_0, \widehat{\beta}_1) &= \arg\min_{(\widehat{\beta}_0,\widehat{\beta}_1)} \sum_{i=1}^n \widehat{\epsilon}_i^2 = \arg\min_{(\widehat{\beta}_0,\widehat{\beta}_1)} \sum_{i=1}^n \left(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)\right)^2 \\ &:= \arg\min_{(\widehat{\beta}_0,\widehat{\beta}_1)} Q(\widehat{\beta}_0, \widehat{\beta}_1).\end{aligned}$$

Theorem 1.2.4 *The least squares estimates are given by*

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)\left(Y_i - \overline{Y}_n\right)}{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}, \tag{1.4}$$

$$\widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n. \tag{1.5}$$

An unbiased estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \widehat{\epsilon}_i^2.$$

Proof. Here we only provide the derivation for the least squares estimates of $\widehat{\beta}_0, \widehat{\beta}_1$ and relegate the derivation for the unbiased estimate of σ^2 to the end of the section. We find the minimum points of $Q(\widehat{\beta}_0, \widehat{\beta}_1)$,

$$\frac{\partial Q}{\partial \widehat{\beta}_0} = 0, \quad \frac{\partial Q}{\partial \widehat{\beta}_1} = 0,$$

to obtain

$$\begin{aligned} \frac{\partial Q}{\partial \widehat{\beta}_0} &= -2\sum_{i=1}^{n}\left(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)\right) = 0, \\ \frac{\partial Q}{\partial \widehat{\beta}_1} &= -2\sum_{i=1}^{n}\left(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)\right) X_i = 0. \end{aligned}$$

Collect the terms to form the normal equation,

$$\begin{aligned} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} X_i &= \sum_{i=1}^{n} Y_i, \\ \widehat{\beta}_0 \sum_{i=1}^{n} X_i + \widehat{\beta}_1 \sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i, \end{aligned} \tag{1.6}$$

to obtain

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)\left(Y_i - \overline{Y}_n\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}, \quad \widehat{\beta}_0 = \overline{Y}_n - \widehat{\beta}_1 \overline{X}_n. \tag{1.7}$$

This must be the minimum point since it is the only critical point of the convex optimization problem. ■

13.6 Example (The 2001 Presidential Election). Figure 13.2 shows the plot of votes for Buchanan (Y) versus votes for Bush (X) in Florida. The least squares estimates (omitting Palm Beach County) and the standard errors are

$$\begin{aligned} \widehat{\beta}_0 &= 66.0991 \quad \widehat{\mathsf{se}}(\widehat{\beta}_0) = 17.2926 \\ \widehat{\beta}_1 &= 0.0035 \quad \widehat{\mathsf{se}}(\widehat{\beta}_1) = 0.0002. \end{aligned}$$

The fitted line is

$$\text{Buchanan} = 66.0991 + 0.0035\,\text{Bush}.$$

(We will see later how the standard errors were computed.) Figure 13.2 also shows the residuals. The inferences from linear regression are most accurate when the residuals behave like random normal numbers. Based on the residual plot, this is not the case in this example. If we repeat the analysis replacing votes with log(votes) we get

$$\begin{aligned} \widehat{\beta}_0 &= -2.3298 \quad \widehat{\mathsf{se}}(\widehat{\beta}_0) = 0.3529 \\ \widehat{\beta}_1 &= 0.730300 \quad \widehat{\mathsf{se}}(\widehat{\beta}_1) = 0.0358. \end{aligned}$$

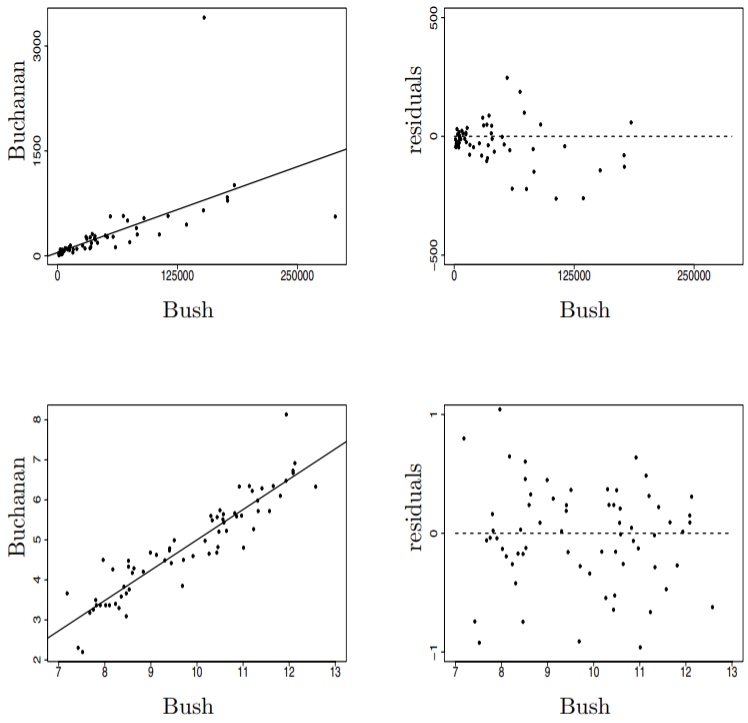


FIGURE 13.2. Voting Data for Election 2000. See example 13.6.

This gives the fit

$$\log(\text{Buchanan}) = -2.3298 + 0.7303\ \log(\text{Bush}).$$

The residuals look much healthier. Later, we shall address the following question: how do we see if Palm Beach County has a statistically plausible outcome? ■

Figure 1.1:

1.3 Least Squares and Maximum Likelihood

Suppose we **add the assumption** that $\epsilon_i|X_i \sim N(0,\sigma^2)$, that is,

$$Y_i|X_i \sim N(\mu_i,\sigma^2)$$

where $\mu_i = \beta_0 + \beta_1 X_i$. The likelihood function is

$$\begin{aligned}\prod_{i=1}^n f(X_i,Y_i) &= \prod_{i=1}^n f_X(X_i) f_{Y|X}(Y_i|X_i) \\ &= \prod_{i=1}^n f_X(X_i) \times \prod_{i=1}^n f_{Y|X}(Y_i|X_i) = L_1 \times L_2,\end{aligned}$$

where $L_1 = \prod_{i=1}^n f_X(X_i)$ and

$$L_2 = \prod_{i=1}^n f_{Y|X}(Y_i|X_i).$$

The term L_1 does not involve the parameters β_0 and β_1. We shall focus on the second term L_2 which is called the **conditional likelihood**, given by

$$L_2 \equiv L(\beta_0,\beta_1,\sigma) = \prod_{i=1}^n f_{Y|X}(Y_i|X_i) \propto \sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\sum_i (Y_i-\mu_i)^2\right\}.$$

The conditional log-likelihood is

$$\ell(\beta_0,\beta_1,\sigma) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i-(\beta_0+\beta_1 X_i))^2$$

We find the MLS estimator,

$$\left(\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\sigma}\right) = \arg\max_{\beta_0,\beta_1,\sigma} \ell(\beta_0,\beta_1,\sigma).$$

For $\widehat{\beta}_0,\widehat{\beta}_1$, we see that maximizing the likelihood is the same as minimizing the RSS.

Theorem 1.3.1 *Under the assumption of Normality, the least squares estima tor is also the maximum likelihood estimator.*

We can also maximize $\ell(\beta_0,\beta_1,\sigma)$ over σ, yielding the MLE

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{\epsilon}_i^2.$$

We take derivative w.r.t. σ^2,

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{(\sigma^2)^2}\sum_{i=1}^n (Y_i-(\beta_0+\beta_1 X_i))^2 = 0, \\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^n (Y_i-(\beta_0+\beta_1 X_i))^2.\end{aligned}$$

Note that MLE estimator $\widehat{\sigma}^2$ is a biased estimator. The unbiased estimator is given by $\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \widehat{\epsilon}_i^2$, which will be proved in the following.

1.4 Properties of the Least Squares Estimators

We now record the standard errors and limiting distribution of the least squares estimator. In regression problems, we usually focus on the proper ties of the estimators conditional on

$$\mathbf{X} = (1, ..., 1; X_1, ..., X_n)^\top.$$

Thus, we can also state the means and variances as conditional means and variances.

Theorem 1.4.1 *Let $\widehat{\boldsymbol{\beta}}^\top = (\widehat{\beta}_0, \widehat{\beta}_1)^\top$ denote the least squares estimators. Then $\widehat{\boldsymbol{\beta}}$ is linear estimator of $Y_1, \ldots, Y_n$ such that*

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}) &= E(\boldsymbol{\beta}|\mathbf{X}) = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \\ Var(\widehat{\boldsymbol{\beta}}) &= Var(\boldsymbol{\beta}|\mathbf{X}) = \frac{\sigma^2}{n s_{XX}^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix} = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}, \end{aligned} \tag{1.8}$$

where the sample variance $s_{XX} = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ and

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix}, \quad \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} = \frac{1}{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2} \begin{pmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{pmatrix}.$$

Example 1.4.2 *Before proving the above theorem, we first write the solution in (1.7) in a compact matrix form. Define*

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \\ Y_n \end{pmatrix}.$$

Then

$$\mathbf{X}^\top \mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix}, \quad \mathbf{X}^\top \mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{pmatrix}.$$

The normal equation (1.6) can be written as

$$\left(\mathbf{X}^\top \mathbf{X}\right) \widehat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y}.$$

Thus, the solution to $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{Y}\right). \tag{1.9}$$

Proof. (1) From (1.4) or (1.9), we see that $\widehat{\boldsymbol{\beta}}$ is a linear estimator of $\mathbf{Y} = (Y_1, \ldots, Y_n)$.
(2) We now see from (1.9) that $\widehat{\boldsymbol{\beta}}$ is unbiased since

$$E(\widehat{\boldsymbol{\beta}}) = E\left[\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{Y}\right)\right] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top E(\mathbf{Y})\right) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}\right) = \boldsymbol{\beta}.$$

(3) We can compute the **covariance matrix** as

$$\begin{aligned} Var(\widehat{\boldsymbol{\beta}}) &= Var(\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{Y}\right)) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top Var(Y)((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top)^\top \\ &= \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \left(\sigma^2 \mathbf{I}\right) \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}. \end{aligned}$$

The other equalities in (1.8) can be easily verified. In general, all components of $\boldsymbol{\beta}$ are not pairwisely independent or pairwisely uncorrelated as can be seen from

$$Var(\boldsymbol{\beta}) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} = \frac{\sigma^2}{n s_{XX}^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix}.$$

Only when $\overline{X}_n = 0$, we have the uncorrelation between $\widehat{\beta}_0$ and $\widehat{\beta}_1$. ■

The estimated standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are obtained by taking the square roots of the corresponding diagonal terms of $Var(\widehat{\boldsymbol{\beta}})$ and inserting the estimate $\widehat{\sigma}$ for σ. Thus,

$$\begin{aligned} \widehat{\sigma}(\widehat{\beta}_0) &= \frac{\widehat{\sigma}}{\sqrt{s_{XX}}\sqrt{n}} \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}, \\ \widehat{\sigma}(\widehat{\beta}_1) &= \frac{\widehat{\sigma}}{\sqrt{s_{XX}}\sqrt{n}}. \end{aligned}$$

We can also write these as $\widehat{\sigma}(\widehat{\beta}_0|\mathbf{X})$ and $\widehat{\sigma}(\widehat{\beta}_1|\mathbf{X})$ but we will use the shorter notation $\widehat{\sigma}(\widehat{\beta}_0)$ and $\widehat{\sigma}(\widehat{\beta}_1)$.

Theorem 1.4.3 *Under appropriate conditions we have:*
1. (Consistency): $\widehat{\beta}_0 \xrightarrow{P} \beta_0$ *and* $\widehat{\beta}_1 \xrightarrow{P} \beta_1$. *(proved using Chebyshev's inequality)*
2. (Asymptotic Normality):

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}(\widehat{\beta}_0)} \xrightarrow{d} N(0,1) \quad \textit{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}(\widehat{\beta}_1)} \xrightarrow{d} N(0,1).$$

3. Approximate $1-\alpha$ *confidence intervals for* β_0 *and* β_1 *are*

$$\widehat{\beta}_0 \pm z_{\alpha/2}\widehat{\sigma}(\widehat{\beta}_0) \quad \textit{and} \quad \widehat{\beta}_1 \pm z_{\alpha/2}\widehat{\sigma}(\widehat{\beta}_1).$$

4. The Wald test for testing $H_0 : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$ *is: reject* H_0 *if* $|W| > z_{\alpha/2}$ *where* $W = (\widehat{\beta}_1 - 0)/\widehat{\sigma}(\widehat{\beta}_1)$. *(Recall that the Wald satistic for testing* $H_0 : \beta_1 = \beta_{1,0}$ *versus* $H_1 : \beta_1 \neq \beta_{1,0}$ *is* $W = (\widehat{\beta}_1 - \beta_{1,0})/\widehat{\sigma}(\widehat{\beta}_1)$*).*

1.5 Hypothesis Test in a Simple Linear Regression

In fact, for any observation data (X_i, Y_i) $(i = 1, 2, \ldots, n)$, one can apply the least squares method to find the regression equation no matter if there is a linear correlation between Y and X. When Y and X are not linearly correlated, it becomes meaningless to compute the linear regression equation. Hence, we need to determine if Y and X are linearly correlated based on our observation data.

If $\beta_1 = 0$, then Y and X are NOT linearly correlated which means that the linear model and the linear regression are not valid. On the other hand, if $\beta_1 \neq 0$ then Y and X are linearly correlated which means that the linear model and the regression are both valid. Thus the hypothesis test is

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0.$$

To test the above hypothesis, we need the following decomposition formula.

Definition 1.5.1 *Define the total sum of squares (TSS) as*

$$\text{TSS} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2. \tag{1.10}$$

The explained sum of squares (ESS) is

$$\text{ESS} = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2. \tag{1.11}$$

The residual sum of squares (RSS) is

$$\text{RSS} = \sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2. \tag{1.12}$$

TSS总偏差平方和，ESS回归平方和，RSS误差平方和

Theorem 1.5.2 *The decomposition formula holds true,*

$$\text{TSS} = \text{ESS} + \text{RSS}.$$

Proof. We compute

$$\begin{aligned}\text{TSS} &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + 2\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2.\end{aligned}$$

The second term vanishes,

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) &= \sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)(\widehat{\beta}_0 + \widehat{\beta}_1 X_i - \overline{Y}) \\ &= \sum_{i=1}^{n}(Y_i - \overline{Y} + \widehat{\beta}_1\overline{X} - \widehat{\beta}_1 X_i)(\overline{Y} - \widehat{\beta}_1\overline{X} + \widehat{\beta}_1 X_i - \overline{Y}) \\ &= \widehat{\beta}_1\sum_{i=1}^{n}(Y_i - \overline{Y} + \widehat{\beta}_1\overline{X} - \widehat{\beta}_1 X_i)(X_i - \overline{X}) \\ &= \widehat{\beta}_1\left[\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X}) + \widehat{\beta}_1\sum_{i=1}^{n}(\overline{X} - X_i)(X_i - \overline{X})\right] \\ &= \widehat{\beta}_1\left[\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X}) - \widehat{\beta}_1\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = 0,\end{aligned}$$

where made use of the definition of $\widehat{\beta}_1$. Thus, the conclusion is verified. ■

From the above, we see that the value of TSS (the sample variance of Y) reveals the diversity of $Y_1, \ldots, Y_n$. The value of ESS reveals the diversity of $\widehat{Y}_1, \ldots, \widehat{Y}_n$ since

$$\text{ESS} = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{\widehat{Y}})^2,$$

where

$$\begin{aligned}\overline{\widehat{Y}} &= \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i = \frac{1}{n}\sum_{i=1}^{n}\widehat{\beta}_0 + \widehat{\beta}_1 X_i = \widehat{\beta}_0 + \widehat{\beta}_1\overline{X} = \overline{Y} - \widehat{\beta}_1\overline{X} + \widehat{\beta}_1\overline{X} \\ &= \overline{Y}.\end{aligned}$$

Moreover, since $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ $(i = 1, \ldots, n)$ all lying on the regression line, the diversity of $\widehat{Y}_1, \ldots, \widehat{Y}_n$ revealed by ESS in fact depends on the diversity of $X_1, \ldots, X_n$. The value of RSS reveals the other factors (such as the noise) which affect the fluctuation of Y besides the factor by linear dependence on X.

The larger ESS corresponding to the smaller RSS will give us a "better" regression equation. Obviously we have

$$0 \leq \frac{\text{ESS}}{\text{TSS}} \leq 1.$$

The following states the relationship between the ratio and the linear relation of Y and X.

ratio	linear dependence relation between Y and X
1	completely linear dependence
close to 1	strongly linear dependence
close to 0	weakly linear dependence
0	completely no linear dependence

Definition 1.5.3 *The correlation between X and Y is defined as*

$$r = \frac{\sum_{i=1}^n X_i Y_i - n\overline{X}\ \overline{Y}}{\sqrt{\sum_{i=1}^n X_i^2 - n\overline{X}^2}\sqrt{\sum_{i=1}^n Y_i^2 - n\overline{Y}^2}},$$

which is a statistic.

Theorem 1.5.4 *There is the following relation among* TSS, ESS, *and the correlation r,*

$$1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\text{ESS}}{\text{TSS}} = r^2, \tag{1.13}$$

where the quantity r^2 is called R-squared.

Proof. We compute

$$\begin{aligned} \text{ESS} &= \sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2 = \sum_{i=1}^n (\widehat{\beta}_0 + \widehat{\beta}_1 X_i - \overline{Y})^2 \\ &= \sum_{i=1}^n (\overline{Y} - \widehat{\beta}_1 \overline{X} + \widehat{\beta}_1 X_i - \overline{Y})^2 = \widehat{\beta}_1^2 \sum_{i=1}^n (X_i - \overline{X})^2, \end{aligned} \tag{1.14}$$

where we see that ESS is a rank 1 quantity. Substituting the expression of $\widehat{\beta}_1$ in (1.7) into above, we arrive at the resulting relation. ■

Using the above theorem, we have $0 \leq r \leq 1$,

the value of r	linear dependence relation between Y and X
$r = 1$	completely linear dependence
r is close to 1	strongly linear dependence
r is close to 0	weakly linear dependence
$r = 0$	completely no linear dependence

Moreover, we can have the following hierarchy,

the value of r	linear dependence relation between Y and X
$r > 0.8$	significantly linear dependence
$0.5 < r \leq 0.8$	strongly linear dependence
$0.3 < r \leq 0.5$	weakly linear dependence
$r \leq 0.3$	nearly no linear dependence

There are several direct testing methods for the validity of linear regressions. The first approach is based on the locations of scattering points. If the points are scattered near one straight line, then the linear regression equation is thought to be valid. The second approach is based on correlation coefficient r. When $r > 0.8$, the linear regression equation is thought to be valid. In the following, we introduce a delicate approach for testing the validity of the linear regression equation. The approach can also be generalized to multivariate linear regression regime. For this testing approach, we need a stronger assumption for the linear regression model.

Definition 1.5.5 *For the linear regression model, $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $(i = 1, \ldots, n)$, if the noises $\{\epsilon_i\}$ are i.i.d.* ***normally distributed*** *with $N(0, \sigma^2)$, then the model is called a normal linear regression model.*

Let the hypothesis test be

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0.$$

We take the statistic

$$F \propto \frac{\text{ESS}}{\text{RSS}}.$$

Based on the result in (1.13), we see that when ESS is large and RSS is small (corresponding to large r^2), there is a significantly linear dependence between Y and X, in which we should reject H_0. Thus, the rejection region is $F \geq C$ for some constant C.

We now derive the distribution of the statistic F. Based on the definition of TSS in (1.10), we see that

$$\frac{\text{TSS}}{\sigma^2} \sim \chi^2(n-1).$$

Based on equation (1.14), we see that $\widehat{\beta}_1$ is normally distributed and ESS has rank of 1 for the quadratic form. For the quadratic form of $Y_1, \ldots, Y_n$,

$$\text{RSS} = \sum_{i=1}^{n} (\widehat{Y}_i - Y_i)^2 = \sum_{i=1}^{n} (\widehat{\beta}_0 + \widehat{\beta}_1 X_i - Y_i)^2,$$

its rank is of $n-2$ since $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are constraint to (1.4) and (1.5). Notice that

$$\frac{\text{TSS}}{\sigma^2} = \frac{\text{ESS}}{\sigma^2} + \frac{\text{RSS}}{\sigma^2}.$$

Since the ranks of $\text{TSS}, \text{ESS}, \text{RSS}$ satisfy $n - 1 = 1 + (n-2)$, we arrive at the result that

$$\frac{\text{ESS}}{\sigma^2} \sim \chi^2(1), \quad \frac{\text{RSS}}{\sigma^2} \sim \chi^2(n-2),$$

and they are independent of each other based on the conclusion of Cochran's Theorem. Hence, we construct the statistic

$$F = \frac{\text{ESS}/1}{\text{RSS}/(n-2)} = (n-2)\frac{\text{ESS}}{\text{RSS}} \sim F(1, n-2),$$

when H_0 is true. We take the significance level α, then the rejection region is

$$F > F_\alpha(1, n-2).$$

Moreover, the statistic F can be computed by

$$F = (n-2)\frac{\text{ESS}}{\text{RSS}} = (n-2)\frac{\text{ESS}}{\text{TSS} - \text{ESS}} = (n-2)\frac{r^2}{1-r^2}.$$

In summary, the validity of a linear regression equation can be tested as follows:
(1) Propose the hypothesis test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.
(2) Compute the statistic $F = (n-2)\frac{r^2}{1-r^2}$.
(3) If $F > F_\alpha(1, n-2)$, then we reject the null H_0 and the linear regression equation is *valid*. If $F \leq F_\alpha(1, n-2)$, then we accept the null H_0 and the linear regression equation is *invalid*.

Example 1.5.6 *In a regression problem for weight Y and height X, the number of samples is* 10 *and the correlation coefficient is $r = 0.91$. We ask whether the linear dependence is significant between Y and X.*
Solution. We compute

$$F = (n-2)\frac{r^2}{1-r^2} = (8)\,\frac{0.91^2}{1-0.91^2} = 37.9 > 5.32 = F_{0.05}(1, 8).$$

Hence we reject the null hypothesis and believe that there is a significantly linear dependence between weight and height.

柯赫伦定理

Theorem 1.5.7 *(Cochran's Theorem) A theorem, given by Cochran in 1934, concerning sum of chi-squared variables. Let Y represent an $n\times 1$ vector of independent standard normal random variables and let $A_1, \ldots, A_k$ be non-zero symmetric matrices such that $\sum_{j=1}^k A_j = I$. Write $Q_j = Y^\top A_j Y$. Cochran's theorem, published in 1934, state that, if any one of the following three conditionis true, then so are the other two.*
(1) The ranks of $A_1, \ldots, A_k$ sum to n which is the rank of Y.
(2) Each of $Q_1, \ldots, Q_k$ has a chi-squared distribution of degrees of freedom of the ranks of $A_1, \ldots, A_k$.
(3) Each of $Q_1, \ldots, Q_k$ is independent of all the others.

1.6 Estimation for the Variance of Noises

The value of σ^2 reflects the well fitness of linear regression. In most cases, σ^2 is unknown so that we need to estimate it. One general idea is to estimate σ^2 by $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \epsilon_i^2$. However, the values of ϵ_i are still not observable. We can estimate them by $\widehat{\epsilon}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$. Therefore,

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{\epsilon}_i^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 = \frac{1}{n}\mathrm{RSS}.$$

However, this estimator is biased and we need to correct it to obtain the unbiased estimator.

Theorem 1.6.1 *For the linear regression model, $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $(i = 1, \ldots, n)$, the noises $\{\epsilon_i\}$ are pairwise uncorrelated and all have the same expected value* 0 *and variance σ^2 (no assumption for the distribution of the noises). Then $\widehat{\sigma}^2 = \frac{1}{n-2}\mathrm{RSS}$ is an unbiased estimator of σ^2.*

Proof. We compute that

$$\begin{aligned} E[(n-2)\,\widehat{\sigma}^2] &= E\left[\mathrm{RSS}\right] = E\left[\mathrm{TSS} - \mathrm{ESS}\right] \\ &= E\left[\sum_{i=1}^n (Y_i - \overline{Y})^2 - \widehat{\beta}_1^2 \sum_{i=1}^n (X_i - \overline{X})^2\right], \end{aligned}$$

where the formula for ESS follows from (1.14). For the first term, we have

$$\begin{aligned} E(Y_i - \overline{Y})^2 &= Var(Y_i - \overline{Y}) + \left[E(Y_i - \overline{Y})\right]^2 \\ &= Var\left[\left(1 - \frac{1}{n}\right) Y_i - \frac{1}{n} \sum_{j=1, j\neq i}^{n} Y_j\right] + \left[\beta_0 + \beta_1 X_i - \beta_0 - \beta_1 \overline{X}\right]^2 \\ &= \left(1 - \frac{1}{n}\right)^2 \sigma^2 + \frac{(n-1)\sigma^2}{n^2} + \beta_1^2 \left(X_i - \overline{X}\right)^2 \\ &= \left(1 - \frac{1}{n}\right) \sigma^2 + \beta_1^2 \left(X_i - \overline{X}\right)^2. \end{aligned}$$

For the second term, we use the following result,

$$Var(\boldsymbol{\beta}) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} = \frac{\sigma^2}{n s_{XX}^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix}.$$

Then

$$\begin{aligned} E\widehat{\beta}_1^2 &= Var(\widehat{\beta}_1) + (E\widehat{\beta}_1)^2 = \frac{\sigma^2}{n s_{XX}^2} + \beta_1^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} + \beta_1^2. \end{aligned}$$

Therefore, we can compute the final results,

$$\begin{aligned} E[(n-2)\,\widehat{\sigma}^2] &= \sum_{i=1}^{n} E(Y_i - \overline{Y})^2 - \left(E\widehat{\beta}_1^2\right) \sum_{i=1}^{n} (X_i - \overline{X})^2 \\ &= \sum_{i=1}^{n} \left[\left(1 - \frac{1}{n}\right) \sigma^2 + \beta_1^2 \left(X_i - \overline{X}\right)^2\right] - \left(\frac{\sigma^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} + \beta_1^2\right) \sum_{i=1}^{n} (X_i - \overline{X})^2 \\ &= (n-1)\sigma^2 + \beta_1^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 - \sigma^2 - \beta_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2 \\ &= (n-2)\sigma^2. \end{aligned}$$

Thus, $E\left[\text{RSS}\right] = (n-2)\sigma^2$ which means that $\widehat{\sigma}^2 = \frac{1}{n-2}\text{RSS}$ is an unbiased estimator of σ^2. ■

1.7 Prediction

Suppose we have estimated a regression model $\widehat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$ from data $(X_1, Y_1), ..., (X_n, Y_n)$. We observe the value $X = x_*$ of the covariate for a new subject and we want to predict their outcome Y_*. An estimate of Y_* is

$$\widehat{Y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_*. \tag{1.15}$$

Using the formula for the variance of the sum of two random variables,

$$\begin{aligned} Var(\widehat{Y}_*) &= Var(\widehat{\beta}_0 + \widehat{\beta}_1 x_*) = Var(\widehat{\beta}_0) + x_*^2 Var(\widehat{\beta}_1) + 2x_* Cov(\widehat{\beta}_0, \widehat{\beta}_1) \\ &= \frac{\sigma^2}{n s_{XX}^2}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 + x_*^2 - 2x_*\overline{X}_n\right) \\ &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\frac{1}{n}\left(\sum_{i=1}^n X_i^2 + nx_*^2 - 2x_*\sum_{i=1}^n X_i\right) \\ &= \frac{\sigma^2 \sum_{i=1}^n (X_i - x_*)^2}{n\sum_{i=1}^n (X_i - \overline{X})^2}. \end{aligned}$$

The estimated standard error $\widehat{\sigma}(\widehat{Y}_*)$ is the square root of this variance, with $\widehat{\sigma}^2$ in place of σ^2. However, the confidence interval for Y_* is **NOT** of the usual form $\widehat{Y}_* \pm z_{\alpha/2}\widehat{\sigma}(\widehat{Y}_*)$. The reason for this is explained in Exercise 10 of Larry book. The correct form of the confidence interval is given in the following theorem.

Theorem 1.7.1 *(Prediction Interval). Under the assumption for the normal distribution for the noises, we have a* $1-\alpha$ *prediction interval for* Y_*,

$$\begin{aligned} & \widehat{Y}_* \pm t_{\alpha/2}(n-2)\widehat{\sigma}\sqrt{1 + \frac{\sum_{i=1}^n (X_i - X_*)^2}{n\sum_{i=1}^n (X_i - \overline{X})^2}} \\ = \; & \widehat{Y}_* \pm t_{\alpha/2}(n-2)\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_* - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}}. \end{aligned}$$

where $\widehat{\sigma}^2 = \frac{1}{n-2}\mathrm{RSS}$ *is unbiased and* $\widehat{Y}_* = \widehat{\beta}_0 + \widehat{\beta}_1 X_*$ *is given in (1.15). If there is no assumption for the normal distribution of the noises, then an approximate* $1-\alpha$ *prediction interval for* Y_* *is*

$$\widehat{Y}_* \pm z_{\alpha/2}\widehat{\sigma}\sqrt{1 + \frac{\sum_{i=1}^n (X_i - X_*)^2}{n\sum_{i=1}^n (X_i - \overline{X})^2}},$$

when the number of data n *is large enough.*

Proof. We know based on the model assumption that

$$\frac{Y - \beta_0 - \beta_1 X}{\sigma} \sim N(0,1).$$

However, β_0, β_1, σ are all unknown in the model so that they need to be replaced. For the denominator, we can use the unbiased $\widehat{\sigma}^2 = \frac{1}{n-2}\mathrm{RSS}$, which is (asymptotically) χ^2 distributed. For the numerator, $Y - \widehat{\beta}_0 - \widehat{\beta}_1 X$ is (asymptotically) normally distributed with mean

$$E(Y - \widehat{\beta}_0 - \widehat{\beta}_1 X) = E(\beta_0 + \beta_1 X + \epsilon - \widehat{\beta}_0 - \widehat{\beta}_1 X) = 0.$$

We can construct the following pivot quantity,

$$\frac{\frac{Y - \widehat{\beta}_0 - \widehat{\beta}_1 X}{\sqrt{Var(Y - \widehat{\beta}_0 - \widehat{\beta}_1 X)}}}{\sqrt{\frac{1}{n-2}\frac{\mathrm{RSS}}{\sigma^2}}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-2)}{n-2}}} \sim t(n-2),$$

where the only unknown is Y. Using the result

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \sigma^2\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} = \frac{\sigma^2}{n s_{XX}^2}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix}\right),$$

we can compute

$$\begin{aligned}
Var(Y-\widehat{\beta}_0-\widehat{\beta}_1X) &= Var(\beta_0-\widehat{\beta}_0+\beta_1X-\widehat{\beta}_1X+\epsilon)\\
&= \sigma^2+Var(\widehat{\beta}_0-\beta_0)+X^2Var(\widehat{\beta}_1-\beta_1)+2XCov(\widehat{\beta}_0-\beta_0,\widehat{\beta}_1-\beta_1)\\
&= \sigma^2\left[1+\frac{1}{n\sum_{i=1}^n(X_i-\overline{X})^2}\left(\sum_{i=1}^n X_i^2+nX^2-2X\sum_{i=1}^n X_i\right)\right]\\
&= \sigma^2\left[1+\frac{\sum_{i=1}^n(X_i-X)^2}{n\sum_{i=1}^n(X_i-\overline{X})^2}\right]\\
&= \sigma^2\left[1+\frac{\sum_{i=1}^n X_i^2-n\overline{X}^2+n\overline{X}^2-2nX\overline{X}+nX^2}{n\sum_{i=1}^n(X_i-\overline{X})^2}\right]\\
&= \sigma^2\left[1+\frac{1}{n}+\frac{(\overline{X}-X)^2}{\sum_{i=1}^n(X_i-\overline{X})^2}\right].
\end{aligned}$$

Thus the pivot quantity can be simplified,

$$\frac{\frac{Y-\widehat{\beta}_0-\widehat{\beta}_1X}{\sqrt{Var(Y-\widehat{\beta}_0-\widehat{\beta}_1X)}}}{\sqrt{\frac{1}{n-2}\frac{\mathrm{RSS}}{\sigma^2}}}=\frac{\frac{Y-\widehat{\beta}_0-\widehat{\beta}_1X}{\sigma\sqrt{1+\frac{\sum_{i=1}^n(X_i-X)^2}{n\sum_{i=1}^n(X_i-\overline{X})^2}}}}{\frac{\widehat{\sigma}}{\sigma}}=\frac{Y-\widehat{\beta}_0-\widehat{\beta}_1X}{\widehat{\sigma}\sqrt{1+\frac{\sum_{i=1}^n(X_i-X)^2}{n\sum_{i=1}^n(X_i-\overline{X})^2}}}.$$

The prediction $1-\alpha$ confidence interval for Y_* at $X=X_*$ is

$$\widehat{Y}_*\pm t_{\alpha/2}(n-2)\widehat{\sigma}\sqrt{1+\frac{\sum_{i=1}^n(X_i-X_*)^2}{n\sum_{i=1}^n(X_i-\overline{X})^2}}.$$

■

1.8 Multiple Regression

1.8.1 Parameter estimation

Now suppose that the covariate is a vector of length k. The data are of the form

$$(Y_1,X_1),...,(Y_i,X_i),...,(Y_n,X_n),$$

where

$$X_i=(X_{i1},...,X_{ik}).$$

Here, X_i is the vector of k covariate values for the ith observation. The linear regression model is

$$Y_i=\sum_{j=1}^k\beta_jX_{ij}+\epsilon_i,$$

for $i=1,...,n$, where $E(\epsilon_i|X_{1i},...,X_{ki})=0$. Usually we want to include an intercept in the model which we can do by **setting** $X_{i1}=1$ **for** $i=1,...,n$. At this point it will be more convenient to express the model in matrix notation. The outcomes will be denoted by

$$\mathbf{Y}=\begin{pmatrix}Y_1\\ \vdots\\ Y_n\end{pmatrix}\in\mathbb{R}^{n\times1},$$

and the covariates will be denoted by

$$\mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nk} \end{pmatrix} \in \mathbb{R}^{n \times k}.$$

Each row is one observation; the columns correspond to the k covariates. Thus, $\mathbf{X}$ is a $(n \times k)$ matrix. Let

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Then we can write the true model as

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

The form of the least squares estimate is given in the following theorem.

Theorem 1.8.1 *Assuming that the $(k \times k)$ matrix $\mathbf{X}^\top\mathbf{X}$ is invertible,*

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}, & (1.16) \\ Var(\widehat{\boldsymbol{\beta}}|\mathbf{X}) &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}, \\ \widehat{\boldsymbol{\beta}} &\approx N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}), & (1.17) \end{aligned}$$

where $\widehat{\boldsymbol{\beta}}$ is a linear unbiased estimator of $\boldsymbol{\beta}$.

The first result can be easily found by

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

Then the solution can be derived by taking the derivative w.r.t. $\boldsymbol{\beta}$. The second and third results can be followed from the previous sections. The estimate regression function is $\widehat{f}(\mathbf{x}) = \sum_{j=1}^k \widehat{\beta}_j x_j$. An unbiased estimate of σ^2 is

$$\widehat{\sigma}^2 = \left(\frac{1}{n-k}\right)\sum_{i=1}^n \widehat{\epsilon}_i^2 = \left(\frac{1}{n-k}\right)\left\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\|_2^2 = \frac{\text{RSS}}{n-k},$$

where $\widehat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ is the vector of residuals. An approximate $1-\alpha$ confidence interval for β_j is

$$\widehat{\beta}_j \pm z_{\alpha/2}\widehat{\sigma}(\widehat{\beta}_j),$$

where $\widehat{\sigma}(\widehat{\beta}_j)$ is the jth diagonal element of the matrix $\widehat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

We now prove that $\widehat{\sigma}^2$ is the unbiased estimate of σ^2.

Theorem 1.8.2 *Assume that $\mathbf{X}$ is full rank with rank of k. $E[\widehat{\sigma}^2] = E[\frac{\text{RSS}}{n-k}] = \sigma^2$.*

Proof. We compute and denote

$$\widehat{\mathbf{E}}_{rr} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = \left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right)\mathbf{Y}.$$

Then

$$\begin{aligned} E\left(\mathrm{RSS}\right) &= E\left\|\widehat{\mathbf{E}}_{rr}\right\|_2^2 = E(\widehat{\mathbf{E}}_{rr}^\top \widehat{\mathbf{E}}_{rr}) = E(\mathrm{tr}[\widehat{\mathbf{E}}_{rr}^\top \widehat{\mathbf{E}}_{rr}]) \\ &= E(\mathrm{tr}[\widehat{\mathbf{E}}_{rr}\widehat{\mathbf{E}}_{rr}^\top]) = \mathrm{tr}(E[\widehat{\mathbf{E}}_{rr}\widehat{\mathbf{E}}_{rr}^\top]). \end{aligned}$$

Since the expected value is zeros,

$$\begin{aligned} E(\widehat{\mathbf{E}}_{rr}) &= E[\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}] = E[\mathbf{Y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}] \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{0}, \end{aligned}$$

then the second order moment can be computed by

$$\begin{aligned} E[\widehat{\mathbf{E}}_{rr}\widehat{\mathbf{E}}_{rr}^\top] &= Var(\widehat{\mathbf{E}}_{rr}) = Var(\left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right)\mathbf{Y}) \\ &= \left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right) Var(\mathbf{Y}) \left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right)^\top \\ &= \left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right)\left(\sigma^2\mathbf{I}\right)\left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right) \\ &= \sigma^2\left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right). \end{aligned}$$

Therefore,

$$\begin{aligned} E\left(\mathrm{RSS}\right) &= \mathrm{tr}(E[\widehat{\mathbf{E}}_{rr}\widehat{\mathbf{E}}_{rr}^\top]) = \sigma^2\mathrm{tr}\left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right) \\ &= \sigma^2\mathrm{tr}\left(\mathbf{I}_n\right) - \mathrm{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right) \\ &= \sigma^2\left(n - \mathrm{tr}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\right) = \sigma^2\left(n-k\right). \end{aligned}$$

■

Theorem 1.8.3 *One has*

$$Cov(\widehat{\mathbf{E}}_{rr}, \widehat{\boldsymbol{\beta}}) = 0.$$

Proof. We compute

$$\begin{aligned} Cov(\widehat{\mathbf{E}}_{rr}, \widehat{\boldsymbol{\beta}}) &= Cov(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}) = Cov(\mathbf{Y}, \widehat{\boldsymbol{\beta}}) - \mathbf{X}Cov(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}) \\ &= Cov(\mathbf{Y}, (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}) - \mathbf{X}Var(\widehat{\boldsymbol{\beta}}) \\ &= Var(\mathbf{Y})[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]^\top - \mathbf{X}\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \sigma^2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} - \sigma^2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = 0. \end{aligned}$$

In above derivations, we only assume the mean and variance of noises but have not assumed the distribution of noises or $\mathbf{Y}$. In the following, we further assume that $\epsilon \sim N(0, \sigma^2)$. ■

Theorem 1.8.4 *Let* $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. *Then (1)* $\widehat{\boldsymbol{\beta}}$ *and* RSS *are independent. (2)* $\mathrm{RSS}/\sigma^2 \sim \chi^2(n-q)$, *where* q *is the rank of the matrix* $\mathbf{X}$.

Proof. (1) Since $\widehat{\mathbf{E}}_{rr}$ and $\widehat{\boldsymbol{\beta}}$ are uncorrelated and they are both normally distributed, they are independent with each other. Since RSS is a function of $\widehat{\mathbf{E}}_{rr}$, then $\widehat{\boldsymbol{\beta}}$ and RSS are independent.

(2) We have the RSS,

$$\begin{aligned} \mathrm{RSS} &= \left[\mathbf{Y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\right]^\top \left[\mathbf{Y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\right] \\ &= \mathbf{Y}^\top\left[\mathbf{I} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]\mathbf{Y}. \end{aligned}$$

We would like to write RSS as the sum of squares of $n-q$ random variables with normal distributions. Let

$$\mathbf{G} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top,$$

which is a symmetric non-negative definite matrix having the same rank with $\mathbf{X}$. Then there exists an orthogonal matrix $\mathbf{C}$ such that

$$\mathbf{CGC}^\top = \begin{pmatrix} \lambda_1 & & \cdots & & 0 \\ & \ddots & & & \\ \vdots & & \lambda_q & & \\ & & & 0 & \vdots \\ & & & & \ddots \\ 0 & & \cdots & & 0 \end{pmatrix}.$$

Since $\mathbf{G}^2 = \mathbf{G}$, thus

$$\mathbf{CGC}^\top = \mathbf{CG}^2\mathbf{C}^\top = \mathbf{CGC}^\top\mathbf{CGC}^\top = \begin{pmatrix} \lambda_1^2 & & \cdots & & 0 \\ & \ddots & & & \\ \vdots & & \lambda_q^2 & & \\ & & & 0 & \vdots \\ & & & & \ddots \\ 0 & & \cdots & & 0 \end{pmatrix}.$$

Therefore,

$$\begin{aligned} \lambda_i^2 &= \lambda_i, \\ \lambda_i &= 1, \quad i = 1, \ldots, q. \\ \mathbf{CGC}^\top &= \begin{pmatrix} \mathbf{I}_q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}. \end{aligned}$$

We take the transformation

$$\mathbf{Z} = \mathbf{C}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right).$$

Then $\mathbf{Z}$ is still normally distributed with

$$\begin{aligned} E(\mathbf{Z}) &= \mathbf{C}E\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right) = 0, \\ Var(\mathbf{Z}) &= \mathbf{C}Var\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)\mathbf{C}^\top = \mathbf{C}\sigma^2\mathbf{I}\mathbf{C}^\top = \sigma^2\mathbf{I}_n. \end{aligned}$$

This means that each component of $\mathbf{Z}$ is independent and normally distributed with $N(0,\sigma^2)$. We compute

$$\begin{aligned}
\text{RSS} &= \mathbf{Y}^\top \left[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]\mathbf{Y} \\
&= (\mathbf{Z}^\top\mathbf{C}+\boldsymbol{\beta}^\top\mathbf{X}^\top)\left[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right](\mathbf{C}^\top\mathbf{Z}+\mathbf{X}\boldsymbol{\beta}) \\
&= \left(\mathbf{Z}^\top\mathbf{C}\left[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]+\boldsymbol{\beta}^\top\mathbf{X}^\top-\boldsymbol{\beta}^\top\mathbf{X}^\top\right)(\mathbf{C}^\top\mathbf{Z}+\mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{Z}^\top\mathbf{C}\left[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right](\mathbf{C}^\top\mathbf{Z}+\mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{Z}^\top\mathbf{C}\left[\mathbf{I}-\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]\mathbf{C}^\top\mathbf{Z} = \mathbf{Z}^\top\mathbf{C}\left[\mathbf{I}-\mathbf{G}\right]\mathbf{C}^\top\mathbf{Z} \\
&= \mathbf{Z}^\top\mathbf{C}\mathbf{C}^\top\mathbf{Z}-\mathbf{Z}^\top\mathbf{C}\mathbf{G}\mathbf{C}^\top\mathbf{Z} = \mathbf{Z}^\top\mathbf{Z}-\mathbf{Z}^\top\begin{pmatrix}\mathbf{I}_q & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{Z} \\
&= z_{q+1}^2+\cdots+z_n^2.
\end{aligned}$$

Thus, RSS is the sum of squares of $n-q$ random variables $(z_{q+1},\cdots,z_n)$ with normal distributions. Thus,

$$\frac{\text{RSS}}{\sigma^2}\sim\chi^2(n-q).$$

■

In the following, we always assume that among the k components, the first one corresponds to the constant term and the others correspond to dimensions of variables. We let

$$k=p+1,$$

so that p is dimension of variables.

Theorem 1.8.5 *Let* ESS *be defined in (1.11). Let* $\mathbf{X}$ *be full rank with rank of* $k=p+1$. *Then*

$$\frac{\text{ESS}}{\sigma^2}\sim\chi^2(p).$$

Proof. Denote $\widehat{\boldsymbol{\beta}}=(\widehat{\beta}_0,\widehat{\beta}_1,\ldots,\widehat{\beta}_p)$. We first write $\widehat{\beta}_0$ in terms of all the other $\widehat{\boldsymbol{\beta}}_{1:p}:=(\widehat{\beta}_1,\ldots,\widehat{\beta}_p)$ in order to show that $\overline{\widehat{Y}}=\overline{Y}$,

$$\widehat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2=\arg\min_{\beta_0,\boldsymbol{\beta}_{1:p}}\|\mathbf{Y}-\mathbf{1}\beta_0-\mathbf{X}_{1:p}\boldsymbol{\beta}_{1:p}\|_2^2.$$

Taking derivative w.r.t. β_0, we obtain

$$\begin{aligned}
-2\left(\mathbf{Y}-\mathbf{1}\widehat{\beta}_0-\mathbf{X}_{1:p}\widehat{\boldsymbol{\beta}}_{1:p}\right)^\top\mathbf{1} &= 0, \\
\sum_{i=1}^n Y_i-\sum_{k=1}^n\sum_{j=1}^p\widehat{\beta}_jX_{kj} &= n\widehat{\beta}_0, \\
\widehat{\beta}_0 &= \overline{Y}-\sum_{j=1}^p\widehat{\beta}_j\overline{X}_{\cdot j}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\overline{\widehat{Y}} &= \frac{1}{n}\sum_{i=1}^n\widehat{Y}_i=\frac{1}{n}\sum_{i=1}^n(\widehat{\beta}_0+\widehat{\beta}_1X_{i1}+\cdots+\widehat{\beta}_pX_{ip}) \\
&= \widehat{\beta}_0+\widehat{\beta}_1\overline{X}_{\cdot 1}+\cdots+\widehat{\beta}_p\overline{X}_{\cdot p}=\overline{Y}-\sum_{j=1}^p\widehat{\beta}_j\overline{X}_{\cdot j}+\widehat{\beta}_1\overline{X}_{\cdot 1}+\cdots+\widehat{\beta}_p\overline{X}_{\cdot p} \\
&= \overline{Y}.
\end{aligned}$$

Then we can compute ESS as,

$$\begin{aligned}\text{ESS} &= \sum_{i=1}^{n}(\widehat{Y}_i-\overline{Y})^2=\sum_{i=1}^{n}(\widehat{Y}_i-\overline{\widehat{Y}})^2=\sum_{i=1}^{n}(\widehat{\beta}_0+\sum_{j=1}^{p}\widehat{\beta}_jX_{ij}-\widehat{\beta}_0-\sum_{j=1}^{p}\widehat{\beta}_j\overline{X}_{\cdot j})^2\\ &= \sum_{i=1}^{n}\left[\sum_{j=1}^{p}(\widehat{\beta}_jX_{ij}-\widehat{\beta}_j\overline{X}_{\cdot j})\right]^2=\sum_{i=1}^{n}\left[\sum_{j=1}^{p}\sum_{k=1}^{p}\widehat{\beta}_j\widehat{\beta}_k(X_{ij}-\overline{X}_{\cdot j})(X_{ik}-\overline{X}_{\cdot k})\right]\\ &= \sum_{j=1}^{p}\sum_{k=1}^{p}\widehat{\beta}_j\widehat{\beta}_k\left[\sum_{i=1}^{n}(X_{ij}-\overline{X}_{\cdot j})(X_{ik}-\overline{X}_{\cdot k})\right]:=\sum_{j=1}^{p}\sum_{k=1}^{p}\widehat{\beta}_j\widehat{\beta}_kA_{jk},\end{aligned}$$

where we see that A_{jk} is the covariance matrix of $\mathbf{X}_{1:p}$, which is symmetric positive definite. Therefore, ESS is the sum of squares of normal random variables $\widehat{\beta}_1,\ldots,\widehat{\beta}_p$ with rank of p. Since the rank of RSS is $n-p-1=n-k$ as proved before, and hence the sum of the rank of RSS and the rank of ESS is

$$n-p-1+p=n-1,$$

which is the same as the rank of

$$\frac{\text{TSS}}{\sigma^2}\sim\chi^2(n-1).$$

We can easily examine the following equality in (1.18). By **Cochran's Theorem**, we conclude that $\frac{\text{ESS}}{\sigma^2}$ and $\frac{\text{RSS}}{\sigma^2}$ are independent with each other, and moreover,

$$\frac{\text{ESS}}{\sigma^2}\sim\chi^2(p),\quad \frac{\text{RSS}}{\sigma^2}\sim\chi^2(n-p-1).$$

■

Theorem 1.8.6 *Let* TSS *be defined in (1.10),* ESS *be defined in (1.11), and* RSS *be defined in (1.12). We still have*

$$\text{TSS}=\text{ESS}+\text{RSS}. \tag{1.18}$$

Theorem 1.8.7 *In summary, let* TSS *be defined in (1.10),* ESS *be defined in (1.11), and* RSS *be defined in (1.12). Let* $\mathbf{X}$ *be full rank with rank of* $k=p+1$*. Then*

$$\frac{\text{TSS}}{\sigma^2}\sim\chi^2(n-1),\quad \frac{\text{ESS}}{\sigma^2}\sim\chi^2(p),\quad \frac{\text{RSS}}{\sigma^2}\sim\chi^2(n-p-1)=\chi^2(n-k).$$

Q: Here I leave one question to the reader. What are the distributions for ESS and RSS if there is a linear dependence among the data X (that is, X is not full rank)?

For the χ^2 distribution, independence assumption is very important. We can numerically and analytically check that $2\chi^2(1)\neq\chi^2(2)$, that is, $p_{2\xi_1^2}(x)\neq p_{\xi_1^2+\xi_2^2}(x)$ for i.i.d. ξ_1 and ξ_2 with standard normal distribution $N(0,1)$.

Example 1.8.8 *Let us derive the centralizing and normalizing regression model. Sometimes we need to first centralize and also normalize the data before constructing the regression model,*

$$Y_i-\overline{Y}=\beta_0+\sum_{j=1}^{p}\beta_j\left(X_{ij}-\overline{X}_{\cdot j}\right)+\epsilon_i,\quad i=1,\ldots,n.$$

Then we follow the formula in (1.16) to estimate the regression coefficients which are similar to those as introduced above.

1.8.2 Hypothesis Test for Multiple Regression

We now focus on hypothesis testing and significance testing problem for the multiple regression. The first problem is if there is a linear dependence relation between Y and $X_1, \ldots, X_p$. If there is no linear relation between them, then all the β_j $(j = 1, \ldots, p)$ should be zero. Then the null hypothesis is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0. \tag{1.19}$$

Based on the above results, $\frac{\text{ESS}}{\sigma^2} \sim \chi^2(p)$, $\frac{\text{RSS}}{\sigma^2} \sim \chi^2(n-p-1)$, we set our testing procedure as follows. When (1.19) is true, we test the hypothesis based on the statistic

$$F = \frac{\text{ESS}/p}{\text{RSS}/(n-p-1)} \sim F(p, n-p-1).$$

Given the significance level α, we reject the null hypothesis (1.19) when $F \geq F_{1-\alpha}(p, n-p-1)$ and then there is a linear dependence relation between Y and $X_1, \ldots, X_p$.

The second problem is if each variate X_j is significant to Y under the condition that Y is linearly dependent on $X_1, \ldots, X_p$. If X_j is not significantly important to Y, then β_j should be zero. Then the null hypothesis is set to be

$$H_0^{(j)} : \beta_j = 0, \quad \text{for } j = 1, \ldots, p. \tag{1.20}$$

Based on the result in (1.17) that $\widehat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2)$, where c_{jj} is the $(j+1)$th diagonal component of $(\mathbf{X}^\top \mathbf{X})^{-1}$ (constant $\mathbf{1}$ vector is included in the first column of $\mathbf{X}$). In addition, $\widehat{\beta}_j$ is independent of $\widehat{\sigma}^2 = \frac{\text{RSS}}{n-p-1}$ based on Theorem 1.8.4. When the null hypothesis (1.20) is true, we can construct the statistic for testing,

$$T_j = \frac{\frac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\sigma}}{\sqrt{\frac{\text{RSS}}{\sigma^2}\frac{1}{n-p-1}}} = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\widehat{\sigma}} = \frac{\widehat{\beta}_j}{\sqrt{c_{jj}}\widehat{\sigma}} \sim t(n-p-1).$$

Given the significance level α, we reject the null hypothesis (1.20) when $|T_j| \geq t_{1-\alpha/2}(n-p-1)$ and then there is a significantly linear dependence relation between Y and X_j. We can repeat the above procedure for all $j = 1, \ldots, p$.

1.9 Bias-Variance Decomposition for Ordinary Least Squares

One can always play with kernel trick to generalize the simple linear regression to the feature space regression. The techniques are all the same but the choices of the features are sometimes tricky.

See references in

(1) Bias_Var_Ridge.pdf,

(2) Benyamin Ghojogh - Elements of Dimensionality Reduction and Manifold Learning,

(3) Learning Theory from First Principles by Francis Bach, etc.

1.9.1 Risk decomposition for OLS

We now go back to Proposition 1.3 to do error analysis for Ordinary Least Squares (OLS) problem. Recall that

$$\mathcal{E}(f) = \int_X (f(x) - f_*(x))^2 d\rho_X + \sigma^2.$$

In our current linear regression setup, we have the following generalization error,

$$\mathcal{E}(\widehat{f}) - \mathcal{E}_* = E\left[\frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}_* - \mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2\right]. \tag{1.21}$$

where $\mathcal{E}_* = \sigma^2$ is the minimum of $\mathcal{E}$, the true model is assumed to be

$$Y_* = f_*(\mathbf{X}) + \epsilon = \beta_{0*} + \beta_{1*}X_1 + \cdots + \beta_{p*}X_p + \epsilon = \mathbf{X}\boldsymbol{\beta}_* + \epsilon,$$

and the estimator of $f_*(\mathbf{X})$ is given by a linear regression function $\widehat{f}(\mathbf{X})$,

$$\widehat{f}(\mathbf{X}) = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_p X_p = \mathbf{X}\widehat{\boldsymbol{\beta}}.$$

The following proposition shows that the minimum can be attained at $\boldsymbol{\beta}_*$, and that is equal to σ^2.

Proposition 1.9.1 *(Risk decomposition for OLS - fixed design). Under the linear model and fixed design assumptions above, for any $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$, we have $\mathcal{E}_* = \sigma^2$ and*

$$\mathcal{E}(\widehat{f}) - \mathcal{E}_* = E\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_{\widehat{\Sigma}}^2,$$

where $\widehat{\Sigma} := \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ is the input covariance matrix and $\|\boldsymbol{\beta}\|_{\widehat{\Sigma}}^2 := \boldsymbol{\beta}^\top\widehat{\Sigma}\boldsymbol{\beta}$. If $\widehat{\boldsymbol{\beta}}$ is now a random variable (such as an estimator of $\boldsymbol{\beta}_*$), then

$$\mathcal{E}(\widehat{f}) - \mathcal{E}_* = \underbrace{\|E[\widehat{\boldsymbol{\beta}}] - \boldsymbol{\beta}_*\|_{\widehat{\Sigma}}^2}_{\text{Bias}} + \underbrace{E[\|\widehat{\boldsymbol{\beta}} - E[\widehat{\boldsymbol{\beta}}]\|_{\widehat{\Sigma}}^2]}_{\text{Variance}}.$$

Proof. We see from equation (1.21) that

$$\begin{aligned}
\mathcal{E}(\widehat{f}) - \mathcal{E}_* &= E[\frac{1}{n}(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_*)^\top(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_*)] \\
&= E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)^\top\frac{1}{n}\mathbf{X}^\top\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)] = E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)^\top\widehat{\Sigma}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)] \\
&= E\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_{\widehat{\Sigma}}^2.
\end{aligned}$$

If $\widehat{\Sigma} := \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ is invertible, then this shows that $\boldsymbol{\beta}_*$ is the unique global minimizer of $\mathcal{E}(\widehat{f})$, and that the minimum value $\mathcal{E}_*$ is equal to σ^2. This shows the first claim.

Now if $\widehat{\boldsymbol{\beta}}$ is random, we perform the usual bias/variance decomposition:

$$\begin{aligned}
\mathcal{E}(\widehat{f}) - \mathcal{E}_* &= E\|\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}}) + E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}_*\|_{\widehat{\Sigma}}^2 \\
&= E\|\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\|_{\widehat{\Sigma}}^2 + 2E\left[\left(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right)\widehat{\Sigma}\left(E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}_*\right)\right] + E\|E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}_*\|_{\widehat{\Sigma}}^2 \\
&= E[\|\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\|_{\widehat{\Sigma}}^2] + \|E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta}_*\|_{\widehat{\Sigma}}^2.
\end{aligned}$$

■

Remark 1.9.2 *The quantity $\|\cdot\|_{\widehat{\Sigma}}$ is called the Mahalanobis distance norm (it is a "true" norm whenever $\widehat{\Sigma}$ is positive definite). It is the norm on the parameter space induced by the input data.*

1.9.2 Statistical Properties of the OLS estimator

We can now analyze the properties of the OLS estimator, which has a closed form $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, with the model $Y = \mathbf{X}\boldsymbol{\beta}_* + \epsilon$. The only randomness comes from ϵ and we thus need to compute expectation of linear and quadratic forms in ϵ. As stated before, the properties of OLS are repeated as follows.

Proposition 1.9.3 *(Estimation properties of OLS). The OLS estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ has the following properties:*
(1) it is unbiased, that is, $E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_*$.
(2) its variance is $Var(\widehat{\boldsymbol{\beta}}) = E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)^\top] = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \frac{\sigma^2}{n}\widehat{\Sigma}^{-1}$; $\widehat{\Sigma}^{-1}$ is often called the precision matrix.

We can now put back the expression of the variance in the risk.

Proposition 1.9.4 *(Risk of OLS). The excess risk of the OLS estimator is equal to*

$$\mathcal{E}(\widehat{f}) - \mathcal{E}_* = \frac{\sigma^2 k}{n},$$

where we assume that $\mathbf{X}$ is full rank of k.

Proof. Note here that the expectation is over ϵ only as we are in the fixed design setting. Using the risk decomposition of Proposition 1.9.1 and the fact that $E[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_*$, we have

$$\mathcal{E}(\widehat{f}) - \mathcal{E}_* = E[\|\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\|_{\widehat{\Sigma}}^2].$$

Then we have

$$\begin{aligned}
\mathcal{E}(\widehat{f}) - \mathcal{E}_* &= \operatorname{tr}\left[E\left(\left[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right]\widehat{\Sigma}\left[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right]\right)\right] \\
&= \operatorname{tr}\left[E\left(\widehat{\Sigma}\left[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right]\left[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right]^\top\right)\right] \\
&= \operatorname{tr}\left[\widehat{\Sigma}E\left(\left[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right]\left[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right]^\top\right)\right] = \operatorname{tr}\left[\widehat{\Sigma}Var\left(\widehat{\boldsymbol{\beta}}\right)\right] \\
&= \operatorname{tr}\left[\widehat{\Sigma}\frac{\sigma^2}{n}\widehat{\Sigma}^{-1}\right] = \frac{\sigma^2 k}{n}.
\end{aligned}$$

■

1.10 Different Model Setups

There are various relations among many machine learning tools like Ordinary Least Squares (OLS), Ridge Linear Regression, Principle Component Analysis (PCA), Independent Component Analysis (ICA), Partial Least Squares (PLS), L_1 regression (see robust_regression.pdf), Quantile Regression (see robust_regression.pdf), etc. Every tool has its own advantage depending on how one uses them.

• Ordinary Least Squares (OLS) is used for regression problem when the covariate $\mathbf{X}$ is full rank.

• Ridge Linear Regression is used for regression when the covariate $\mathbf{X}$ is high dimensional and $\mathbf{X}$ is NOT but close to full rank (there are linear dependences among dimensions of $\mathbf{X}$). In my view, ridge regression is good for the case that the number of feature is less than but close to the number of regression coefficients.

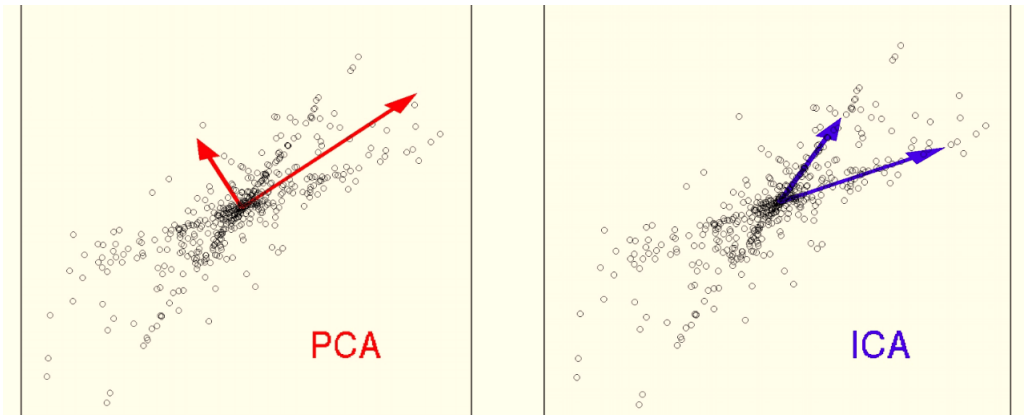


Figure 1.2: PCA vs. ICA.

• LASSO regression stands for Least Absolute Shrinkage and Selection Operator, which is used when the number of feature is much less than the number of regression coefficients (the rank of $\mathbf{X}$ is much smaller than the full rank).

• Principle Component Analysis (PCA) is used for dimension reduction and principle **orthogonal** component detection.

• Independent Component Analysis (ICA) is used for separating different types of signals (see Fig. 1.2).

• The idea behind Principal Component Regression (PCR) is to first perform a principal component analysis (PCA) on the design matrix $\mathbf{X}$ and then use only the first principal components to do the regression.

• Partial Least Squares (PLS) combines PCA and multiple regression to regress when $\mathbf{X}$ is far away from full rank or very low rank. (see PLS_simple_explanation.pdf) The idea behind PLS is to decompose both the design matrix $\mathbf{X}$ and response matrix $\mathbf{Y}$ (the general case of multiple responses is often considered) like in principle component analysis.

Here I only know a little about these methodologies and I only study them a little bit by myself.

See references in my local computers file folders, 2018.04.22 DM_ICA_PCA, 2018.05.29 REU Program, 2018.06.04 Regression. huiguifenxi regression.

See references in my local computers, OLS_OR_MLE_PCA.pdf, linear regression model two noises 76-1-141.pdf, OLS_PCA.pdf, PLS_simple_explanation.pdf, lasso high-dimensional regression.pdf, 12.Robust.pdf, PLS-pretty-Abdi.pdf, robust_regression.pdf, Intro_to_PCA_and_ICA.pdf, Robustness_Multivariate_Orthogonal.pdf, PCA_ICA_compare.pdf.

See website on What is LASSO Regression Definition, Examples and Techniques.html, Lasso regression — Introduction to Regression Models.html, https://stat151a.berkeley.edu/spring-2024/lectures/Lecture23.html (local is [Good] Lasso or 'L1' regression.html), Visually differentiating PCA and Linear Regression _ Know Thy Data.html.

See codes in regression.mw.

See more in my original hand-writing notes.

1.10.1 Both Variables Have Errors

Suppose both X and Y contain some random errors, ϵ_X and ϵ_Y, which may come from measurement or other resources. A suitable model is as follows,

$$\begin{aligned} X &= \xi + \epsilon_X, \qquad \epsilon_X \sim N(0, \sigma_X^2), \\ Y &= \alpha + \beta\xi + \epsilon_Y, \quad \epsilon_Y \sim N(0, \sigma_Y^2), \end{aligned}$$

where ϵ_X and ϵ_Y are independent random measurement errors. There are two analysis approaches concerning this model: the functional and the structural. The basic difference between the two approaches is whether to consider ξ as a non-random variable or a random variable following normal distribution with mean μ and variance τ^2,

$$\xi \sim N(\mu, \tau^2),$$

and independent to both random errors. Since the latter approach is more general, in the discussion below, we will follow the structural model where X and Y follow a bivariate normal distribution with mean and covariance structure as follows:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \alpha + \beta\mu \end{pmatrix}, \begin{pmatrix} \tau^2 + \sigma_X^2 & \beta\tau^2 \\ \beta\tau^2 & \beta^2\tau^2 + \sigma_Y^2 \end{pmatrix} \right),$$

where

$$\begin{aligned} Cov(X, Y) &= (EXY) - (EX)(EY) = E(\alpha\xi + \beta\xi^2) - \mu\,(\alpha + \beta\mu) \\ &= \alpha\mu + \beta\left(\tau^2 + \mu^2\right) - \mu\,(\alpha + \beta\mu) = \beta\tau^2. \end{aligned}$$

Given a random sample of observed X's and Y's, we can obtain the MLE of the slope of the regression. Its value, however, depends on the ratio of the two error variances

$$\gamma = \sigma_Y^2 / \sigma_X^2,$$

to have

$$\widehat{\beta} = \frac{S_{YY} - \gamma S_{XX} + \sqrt{(S_{YY} - \gamma S_{XX})^2 + 4\gamma S_{XY}^2}}{2S_{XY}},$$

where

$$S_{XX} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \quad S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \quad S_{YY} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

We now derive the MLE of β. The computation is extremely complicated so I used Maple for help. Denote

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(m, \Sigma),$$

where

$$m = \begin{pmatrix} \mu \\ \alpha + \beta\mu \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \tau^2 + \sigma_X^2 & \beta\tau^2 \\ \beta\tau^2 & \beta^2\tau^2 + \sigma_Y^2 \end{pmatrix}.$$

The likelihood for one data is

$$p(X, Y | \alpha, \beta, \mu, \tau, \sigma_X^2, \sigma_Y^2) = \frac{1}{2\pi\left|\Sigma\right|^{1/2}} \exp\left(-\frac{1}{2}(z - m)^\top \Sigma^{-1}(z - m) \right).$$

Since (x_i, y_i) are i.i.d., the log likelihood for all the data is

$$\ln \prod_{i=1}^{n} p(x_i, y_i) = -n \ln(2\pi) - \frac{n}{2} \ln |\Sigma| - \sum_{i=1}^{n} \frac{1}{2} (z_i - m)^\top \Sigma^{-1} (z_i - m),$$

where

$$\begin{aligned} \Sigma^{-1} &= \frac{1}{|\Sigma|} \begin{pmatrix} \beta^2\tau^2 + \sigma_Y^2 & -\beta\tau^2 \\ -\beta\tau^2 & \tau^2 + \sigma_X^2 \end{pmatrix}, \\ |\Sigma| &= \beta^2\tau^2\sigma_X^2 + \tau^2\sigma_Y^2 + \sigma_X^2\sigma_Y^2. \end{aligned}$$

We first assume σ_X^2 and σ_Y^2 are given and denote $\gamma = \sigma_Y^2/\sigma_X^2$ in order to find α, β, μ, τ as functions of σ_X^2 and σ_Y^2 (or equivalently σ_X^2 and γ). We change the variables

$$\lambda_0 = \mu, \quad \lambda_1 = \alpha + \beta\mu, \quad \lambda_2 = \beta^2\tau^2 + \gamma\tau^2 + \gamma\sigma_X^2, \quad \beta = \beta,$$

to obtain that

$$\begin{aligned} |\Sigma| &= \sigma_X^2 \left(\beta^2\tau^2 + \gamma\tau^2 + \gamma\sigma_X^2\right) = \lambda_2\sigma_X^2, \\ \Sigma^{-1} &= \frac{1}{|\Sigma|} \begin{pmatrix} \beta^2\tau^2 + \sigma_Y^2 & -\beta\tau^2 \\ -\beta\tau^2 & \tau^2 + \sigma_X^2 \end{pmatrix} = \frac{1}{\lambda_2\sigma_X^2(\beta^2+\gamma)} \begin{pmatrix} \beta^2\lambda_2 + \gamma^2\sigma_X^2 & -\beta(\lambda_2 - \gamma\sigma_X^2) \\ -\beta(\lambda_2 - \gamma\sigma_X^2) & \lambda_2 + \beta^2\sigma_X^2 \end{pmatrix}. \end{aligned}$$

We write the log likelihood function in the new parameterization as

$$\begin{aligned} & l(\lambda_0, \lambda_1, \lambda_2, \beta | x_i, y_i, \sigma_X^2, \sigma_Y^2) \\ = \; & -n \ln(2\pi) - \frac{n}{2} \ln \sigma_X^2 - \frac{n}{2} \ln \lambda_2 \\ & - \left[\left(\beta^2\lambda_2 + \gamma^2\sigma_X^2\right) \sum_{i=1}^{n} (x_i - \lambda_0)^2 - 2\beta(\lambda_2 - \gamma\sigma_X^2) \sum_{i=1}^{n} (x_i - \lambda_0)(y_i - \lambda_1) \right. \\ & \left. + \left(\lambda_2 + \beta^2\sigma_X^2\right) \sum_{i=1}^{n} (y_i - \lambda_1)^2 \right] \Big/ \left(2\lambda_2\sigma_X^2(\beta^2 + \gamma)\right). \end{aligned} \tag{1.22}$$

First we compute λ_0 and λ_1,

$$\begin{aligned} \frac{\partial l}{\partial \lambda_0} &= -\frac{1}{2\lambda_2\sigma_X^2(\beta^2+\gamma)} \left[\left(\beta^2\lambda_2 + \gamma^2\sigma_X^2\right) \sum_{i=1}^{n} 2(\lambda_0 - x_i) + 2\beta(\lambda_2 - \gamma\sigma_X^2) \sum_{i=1}^{n} (y_i - \lambda_1) \right] = 0, \\ \frac{\partial l}{\partial \lambda_1} &= -\frac{1}{2\lambda_2\sigma_X^2(\beta^2+\gamma)} \left[2\beta(\lambda_2 - \gamma\sigma_X^2) \sum_{i=1}^{n} (x_i - \lambda_0) + \left(\lambda_2 + \beta^2\sigma_X^2\right) \sum_{i=1}^{n} 2(\lambda_1 - y_i) \right] = 0. \end{aligned}$$

Thus we can easily observe that

$$\lambda_0 = \overline{x}, \quad \lambda_1 = \overline{y}.$$

Substituting above back into (1.22) and replacing with S_{XX}, S_{XY}, S_{YY}, we obtain the log likelihood,

$$\begin{aligned} & l(\lambda_2, \beta | x_i, y_i, \sigma_X^2, \sigma_Y^2, \lambda_0, \lambda_1) \\ = \; & -n \ln(2\pi) - \frac{n}{2} \ln \sigma_X^2 - \frac{n}{2} \ln \lambda_2 \\ & - \left[nS_{XX} \left(\beta^2\lambda_2 + \gamma^2\sigma_X^2\right) - 2\beta(\lambda_2 - \gamma\sigma_X^2) nS_{XY} + \left(\lambda_2 + \beta^2\sigma_X^2\right) nS_{YY} \right] / \left(2\lambda_2\sigma_X^2(\beta^2 + \gamma)\right). \end{aligned}$$

Taking derivative w.r.t. λ_2,

$$\begin{aligned}\frac{\partial l}{\partial \lambda_2} &= \frac{n\left(S_{XX}\gamma^2 + 2S_{XY}\beta\gamma + S_{YY}\beta^2 - \beta^2\lambda_2 - \gamma\lambda_2\right)}{2\lambda_2^2(\beta^2+\gamma)} = 0,\\ \lambda_2 &= \frac{S_{XX}\gamma^2 + 2S_{XY}\beta\gamma + S_{YY}\beta^2}{\beta^2+\gamma}.\end{aligned}$$

Taking derivative w.r.t. β,

$$\begin{aligned}\frac{\partial l}{\partial \beta} &= \left[S_{XX}\beta\gamma^2\sigma_X^2 + S_{XY}\beta^2\gamma\sigma_X^2 - S_{XY}\gamma^2\sigma_X^2 - S_{YY}\beta\gamma\sigma_X^2\right.\\ &\quad \left. -S_{XX}\beta\gamma\lambda_2 - S_{XX}\beta^2\lambda_2 + S_{XY}\gamma\lambda_2 + S_{YY}\beta\lambda_2\right] \times \frac{n}{\lambda_2\sigma_X^2(\beta^2+\gamma)^2} = 0.\end{aligned}$$

Substituting λ_2, then we obtain

$$-\frac{\left(S_{XX}\beta\gamma + S_{XY}\beta^2 - S_{XY}\gamma - S_{YY}\beta\right)\left(-\beta^2\gamma\sigma_X^2 - \gamma^2\sigma_X^2 + S_{XX}\gamma^2 + 2S_{XY}\beta\gamma + S_{YY}\beta^2\right)}{(\beta^2+\gamma)} = 0.$$

Numerically, only the following solution makes sense,

$$\begin{aligned}&S_{XY}\beta^2 + S_{XX}\beta\gamma - S_{YY}\beta - S_{XY}\gamma = 0,\\ &\widehat{\beta} = \frac{S_{YY} - \gamma S_{XX} + \sqrt{(S_{YY}-\gamma S_{XX})^2 + 4\gamma S_{XY}^2}}{2S_{XY}}.\end{aligned} \tag{1.23}$$

One possibly further get estimators for $\alpha, \beta, \mu, \tau(\sigma_X^2, \sigma_Y^2)$ from $\lambda_0, \lambda_1, \lambda_2, \beta$. Finally one can get estimators for σ_X^2, σ_Y^2 or σ_X^2, γ from the likelihood.

Inference (hypothesis test, confidence interval) on the slope parameter can be carried out similarly using the maximum likelihood approach. We consider this the general and correct approach when both variables are random. Since it is a parametric model, the readers are reminded that normality transformation should be performed prior to the regression analysis if a variable is found not normal.

1.10.2 Ordinary Least Squares (OLS) Regression

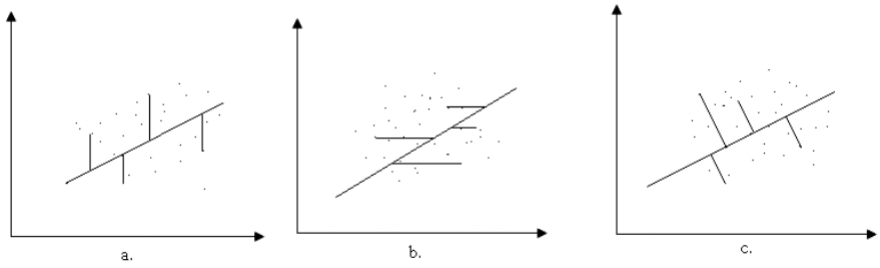


Figure 1.3: The OLS regression (a) and (b) and the OR (c).

As illustrated in Figure 1.3(a), the ordinary least square (OLS) estimate of Y on X will minimize the squared vertical distance $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ from the points to the regression line. The OLS estimate of the slope is $\widehat{\beta} = S_{XY}/S_{XX}$. This is the case when $\gamma = \infty$ in the general structural modelling approach (equation (1.23)). Similarly, the OLS estimate of X on Y would minimize the horizontal distance to the regression line $\sum_{i=1}^{n}(x_i - \alpha - \beta y_i)^2$ (see Fig. 1.3(b)). The OLS estimate of the slope is $\widehat{\beta} = S_{XY}/S_{YY}$ which corresponds the inverse of the result (1.23) when $\gamma = 0$. The latter is also called the reverse regression. Notice that the OLS is suitable when only one of the two variables is random.

1.10.3 Orthogonal Regression (OR)

Instead of minimizing the vertical (or horizontal) distance as in the OLS, the orthogonal regression (OR) takes the middle ground by minimizing the orthogonal distance from the observed data points to the regression line as illustrated in Figure 1.3(c). The resulting OR estimate of β is:

$$\widehat{\beta} = \frac{S_{YY} - S_{XX} + \sqrt{(S_{YY} - S_{XX})^2 + 4S_{XY}^2}}{2S_{XY}}.$$

This is the same as the MLE in the general structural modelling approach when $\gamma = 1$. It means that the orthogonal regression is suitable when the error variances are equal. Let us now minimize the orthogonal distance to the fitted line, $y - \alpha - \beta x = 0$,

$$\min_{\alpha,\beta} l(\alpha, \beta) := \sum_{i=1}^{n} \left(\frac{|y_i - \alpha - \beta x_i|}{\sqrt{\beta^2 + 1}} \right)^2 = \sum_{i=1}^{n} \frac{(y_i - \alpha - \beta x_i)^2}{\beta^2 + 1}.$$

Taking derivative w.r.t. α, $\partial l / \partial \alpha = 0$, we obtain that

$$\alpha = \overline{y} - \beta \overline{x}.$$

Taking derivative w.r.t. β,

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{2(\alpha + \beta x_i - y_i)x_i}{\beta^2 + 1} - \frac{(y_i - \alpha - \beta x_i)^2}{(\beta^2 + 1)^2} 2\beta \right] = 0.$$

We simplify above,

$$\begin{aligned} &\left(\beta^2 + 1\right) \left(\beta \sum x_i^2 + \alpha \sum x_i - \sum x_i y_i \right) \\ &-\beta \left(\sum y_i^2 + \beta^2 \sum x_i^2 + n\alpha^2 - 2\alpha \sum y_i - 2\beta \sum x_i y_i + 2\alpha\beta \sum x_i \right) = 0. \end{aligned}$$

Using Maple to substitute in $\alpha = \overline{y} - \beta \overline{x}$, and also denoting $D_{XX} = \sum x_i^2, D_{XY} = \sum x_i y_i, D_{YY} = \sum y_i^2$,

$$\begin{aligned} \left(D_{XY} - n\overline{xy}\right) \beta^2 + (D_{XX} - D_{YY} - n\overline{x}^2 + n\overline{y}^2)\beta + n\overline{xy} - D_{XY} &= 0, \\ S_{XY}\beta^2 + (S_{XX} - S_{YY})\beta - S_{XY} &= 0. \end{aligned}$$

We finally obtain

$$\widehat{\beta} = \frac{S_{YY} - S_{XX} + \sqrt{(S_{YY} - S_{XX})^2 + 4S_{XY}^2}}{2S_{XY}}.$$

1.10.4 The Connection Between OR and PCA

There is a close relationship between the Principle Component Analysis (PCA) and the Orthogonal Regression. For the sample covariance matrix of the random variables (X, Y), $[S_{XX}, S_{XY}; S_{XY}, S_{YY}]$, its highest eigenvalue (or equivalently the SVD of the **centralized** data) is

$$\begin{aligned} (S_{XX} - \widehat{\lambda})(S_{YY} - \widehat{\lambda}) - S_{XY}^2 &= 0, \\ \widehat{\lambda}^2 - (S_{XX} + S_{YY})\widehat{\lambda} + S_{XX}S_{YY} - S_{XY}^2 &= 0. \end{aligned}$$

$$\widehat{\lambda} = \frac{S_{XX} + S_{YY} + \sqrt{(S_{XX} + S_{YY})^2 - 4(S_{XX}S_{YY} - S_{XY}^2)}}{2}.$$

And the eigenvector (first principal component) corresponding to this eigenvalue is

$$\left(S_{XY}, \frac{S_{YY} - S_{XX} + \sqrt{(S_{YY} - S_{XX})^2 + 4S_{XY}^2}}{2}\right).$$

Therefore, the slope of the first principal component is

$$\widehat{\beta} = \frac{S_{YY} - S_{XX} + \sqrt{(S_{YY} - S_{XX})^2 + 4S_{XY}^2}}{2S_{XY}},$$

which is the same as the slope estimator from the orthogonal regression.

Intuitively, the first principal component is the line passing through the greatest dimension of the concentration ellipse, which coincides with the orthogonal regression line. Therefore, existing statistical inference techniques for the PCA can be applied directly to the inference of the slope parameter from the OR approach

1.11 Introduction to Comparison between Ridge Regression and Lasso Regression

OLS is not robust to outliers. It can produce misleading results if unusual cases go undetected — even a single case can have a significant impact on the fit of the regression surface. We first define the **canonical regularizers**: ℓ_0, ℓ_1, ℓ_2. In regression, arguably the three canonical choices for regularizers are the ℓ_0, ℓ_1, ℓ_2 norms:

$$\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{k} 1\{\beta_j \neq 0\}, \quad \|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{k} |\beta_i|, \quad \|\boldsymbol{\beta}\|_2 = \left(\sum_{i=1}^{k} \beta_i^2\right)^{1/2}.$$

Critically, $\|\cdot\|_0$ is not convex, while $\|\cdot\|_1$ and $\|\cdot\|_2$ are convex. This makes best subset selection a nonconvex problem, and one that is generally very hard to solve in practice except for very small k (dimension of parameters). On the other hand, the lasso and ridge regression problems are convex, and many efficient algorithms exist for them.

1.11.1 Ridge Regression

Experimental and theoretical studies show that PLS (see PLS_simple_explanation.pdf), Principal Component Regression (PCR) (see PLS_simple_explanation.pdf), and ridge regression tend to behave similarly. Ridge regression maybe preferred for its relative interpretational and computational simplicity for low dimensional paramaters.

Ridge regression is a popular form of regularised linear regression, in which we change the objective function from the standard least squares formulation to the following,

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2,$$

for a given value of λ. The solution can be shown to be

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

The "right" value of λ for a given problem is usually obtained through **cross validation**. One problem with this might be that the solution $\widehat{\boldsymbol{\beta}}_{\text{ridge}}$ is still "dense", meaning that, in general, every entry of it is nonzero, and

we still have to invert a dense $k \times k$ matrix. **In my opinion, ridge regression is good for a degenerate but close to a full rank $\mathbf{X}^\top\mathbf{X}$ matrix whereas it is not good for a very low rank $\mathbf{X}^\top\mathbf{X}$ matrix when the dimension of parameters k is large**.

For example, consider our highly correlated regressor example. The ridge regression will still include both regressors, and their coefficient estimates will still be highly negatively correlated, but both will be shrunk towards zero. Maybe it would make more sense to select only one variable to include. Let us try to think of how we can change the penalty term to achieve this.

A "sparse" solution is an estimator $\widehat{\boldsymbol{\beta}}$ in which many of the entries are zero — that is, an estimated regression line that does not use many of the available regressors. In a word — **ridge regression estimates are not sparse**. Let's try to derive one that is by changing the penalty. A very intuitive way to produce a sparse estimate is as follows:

$$\min_{\boldsymbol{\beta}} \left(\frac{1}{n} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_0 \right) = \left(\frac{1}{n} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \sum_{j=1}^{k} 1\{\beta_j \neq 0\} \right),$$

however, this is **practically difficult since the problem is nonconvex**. This finds a tradeoff between the best fit to the data, but with a penalty for using more regressors. This makes sense, but is very difficult to compute. In particular, this objective is very non-convex. Bayesian statisticians do attempt to estimate models with a similar kind of penalty (they are called "spike and slab" models), but they are extremeley computationally intensive and beyond the scope of this course.

1.11.2 Lasso Regression

A convex approximation to the preceding loss is the L^1 or Lasso loss, leading to Lasso or L^1 regression. The popular form of regularised linear regression is lasso, which solves the following problem:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1,$$

where $\left\| \boldsymbol{\beta} \right\|_1 = \sum_i |\beta_i|$. This loss is convex (beacuse it is the sum of two convex functions), and so is much easier to minimize. Furthermore, **as λ grows, it does produce sparser and sparser solutions** — though it may not be obvious at first.

What is Lasso Regression?

LASSO regression, also known as L^1 regularization, is a popular technique used in statistical modeling and machine learning to estimate the relationships between variables and make predictions. LASSO stands for Least Absolute Shrinkage and Selection Operator. The primary goal of LASSO regression is to find a balance between model simplicity and accuracy. It achieves this by adding a penalty term to the traditional linear regression model, which encourages sparse solutions where some coefficients are forced to be exactly zero. This feature makes LASSO particularly useful for **feature selection**, as it can automatically identify and discard irrelevant or redundant variables.

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. Lasso

Regression uses L^1 regularization technique (will be discussed later in this article). It is used when we have more features because it automatically performs feature selection.

L^1 Regularization

Regularization is an important concept that is used to avoid overfitting of the data, especially when the trained and test data are much varying. Regularization is implemented by adding a "penalty" term to the best fit derived from the trained data, to achieve a lesser variance with the tested data and also restricts the influence of predictor variables over the output variable by compressing their coefficients. In regularization, what we do is normally we keep the same number of features but reduce the magnitude of the coefficients. We can reduce the magnitude of the coefficients by using different types of regression techniques which uses regularization to overcome this problem. So, let us discuss them.

LASSO regression introduces an additional penalty term based on the absolute values of the coefficients. The L^1 regularization term is the sum of the absolute values of the coefficients multiplied by a tuning parameter λ:

$$L^1 = \lambda \sum_i |\beta_i|$$

where λ is the regularization parameter that controls the amount of regularization applied and β_i $(i = 1, \ldots, k)$ are the regression coefficients.

Shrinking Coefficients

By adding the L^1 regularization term, LASSO regression can shrink the coefficients towards zero. When λ is sufficiently large, some coefficients are driven to exactly zero. This property of LASSO makes it useful for feature selection, as the variables with zero coefficients are effectively removed from the model.

Tuning parameter λ

The choice of the regularization parameter λ is crucial in LASSO regression. A larger λ value increases the amount of regularization, leading to more coefficients being pushed towards zero. Conversely, a smaller λ value reduces the regularization effect, allowing more variables to have non-zero coefficients.

- λ denotes the amount of shrinkage.
- $\lambda = 0$ implies all features are considered and it is equivalent to the linear regression where only the residual sum of squares is considered to build a predictive model
- $\lambda = \infty$ implies no feature is considered i.e, as λ closes to infinity it eliminates more and more features
- The bias increases with increase in λ
- The variance increases with decrease in λ

Model Fitting

To estimate the coefficients in LASSO regression, an optimization algorithm is used to minimize the objective function. **Coordinate Descent** is commonly employed, which iteratively updates each coefficient while holding the others fixed.

By striking a balance between simplicity and accuracy, LASSO can provide interpretable models while effectively managing the risk of overfitting. It's worth noting that LASSO is just one type of regularization technique, and there are other variants such as Ridge regression (L^2 regularization) and Elastic Net.

Lasso Meaning

LASSO regression offers a powerful framework for both prediction and feature selection, especially when dealing with high-dimensional datasets where the number of features is large. The word "LASSO" stands for

Least Absolute Shrinkage and Selection Operator. It is a statistical formula for the regularisation of data models and feature selection.

Standardization

Lasso performs best when all numerical features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

This means it is important to standardize our features. We do this by subtracting the mean from our observations and then dividing the standard deviation. This so called standard score Z for an observation X is calculated as:

$$Z = \frac{X - \overline{X}}{s},$$

where $X = (X_1, \ldots, X_n)$ is an observation in one feature, $\overline{X}$ is the mean of that feature, and s is the standard deviation of that feature.

The Lasso Produces Sparse Solutions (Intuition)

One way to see that the Lasso produces sparse solutions is to start with a very large λ and see what happens as it is slowly decreased.

Start at λ very large, so that $\widehat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda) = 0$. If we take small step of size ε in a particular direction away from zero in entry β_j, then $\lambda\|\widehat{\boldsymbol{\beta}}\|_1$ increases by $\varepsilon\lambda$, and the RSS changes by the gradient of the squared error,

$$\varepsilon \sum_{i=1}^{n} (y_i - \widehat{\boldsymbol{\beta}}(\lambda)\mathbf{x}_{i\cdot}) x_{ij} := \varepsilon \sum_{i=1}^{n} \widehat{\varepsilon}_i x_{ij} = \varepsilon \sum_{i=1}^{n} y_i x_{ij}, \quad (\text{because } \widehat{\boldsymbol{\beta}}(\lambda) = 0).$$

As long as $|\sum_{i=1}^{n} y_i x_{ij}| < \lambda$ for all $j \in \{1, \ldots, k\}$, we cannot improve the loss by moving away from 0. Since the loss is convex, that means 0 is the minimum.

Eventually, we decrease λ until $\sum_{i=1}^{n} y_i x_{ij} = \lambda$ for some j (**greedy variable selection**). At that point, β_j moves away from zero as λ decreases, and the $\widehat{\varepsilon}_i$ also change. However, until $\sum_{i=1}^{n} \widehat{\varepsilon}_i x_{iq} = \lambda$ for some other $q \neq j$, only β_j will be nonzero. As λ decreases more and more, variables tend to get added to the model, until $\lambda = 0$, when of course $\widehat{\boldsymbol{\beta}}_{\text{lasso}}(0) = \widehat{\boldsymbol{\beta}}_{\text{OLS}}$, the OLS solution.

Conclusion

LASSO regression emerges as a crucial technique for statistical modeling and machine learning, striking a balance between model simplicity and accuracy.

With its ability to promote sparsity through feature selection, LASSO regression aids in identifying relevant variables and managing overfitting, particularly in high-dimensional datasets.

See more details about Lasso regression in Learning from First Principles by Bach.

1.11.3 Comparison in Short

In short, Ridge is a shrinkage model, and Lasso is a feature selection model. Ridge tries to balance the bias-variance trade-off by shrinking the coefficients, but it does not select any feature and keeps all of them. Lasso tries to balance the bias-variance trade-off by shrinking some coefficients to zero. In this way, **Lasso can be seen as an optimizer for feature selection**. See Table 1.1 for more comparisons. Also see Fig. 1.4 for illustration.

Table 1.1: Comparison between Ridge Regression and LASSO Regression.

	Ridge Regression	LASSO Regression
Penalty Term	The penalty term is the sum of the squares of the coefficients (L^2 regularization).	The penalty term is the sum of the absolute values of the coefficients (L^1 regularization).
Shrinkage	Shrinks the coefficients but does not set any coefficient to zero.	Can shrink some coefficients to zero, effectively performing feature selection.
Overfitting	Helps to reduce overfitting by shrinking large coefficients.	Helps to reduce overfitting by shrinking and selecting features with less importance.
Number of Features	Works well when there are a large number of features.	Works well when there are a small number of features.
Thresholding	Performs "soft thresholding" of coefficients.	Performs "hard thresholding" of coefficients.
Convexity	Always strictly convex. We are guaranteed a unique ridge solution.	Not strictly convex when $k > n$. We are not necessarily to have a unique Lasso solution.

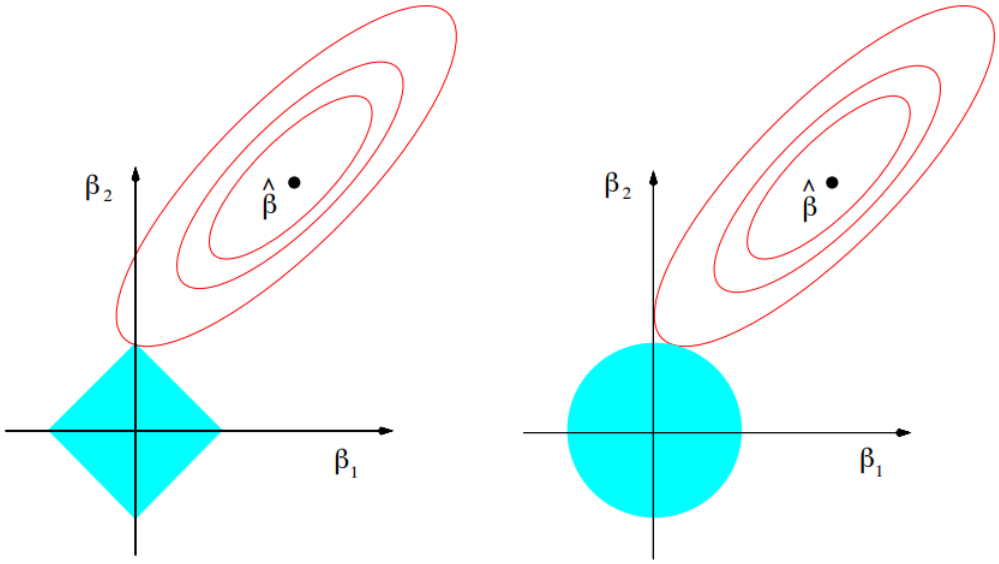


Figure 1.4: The "classical" illustration comparing lasso and ridge constraints. See Chap. 3.4 of Hastie et al. (2009).

1.12 Bias-Variance Tradeoff in Ridge Linear Regression

1.12.1 Least-squares in high dimensions.

When k/n approaches 1, we are essentially memorizing the observations y_i (that is, for example when $k = n$ and $\mathbf{X}$ is a square invertible matrix, $\boldsymbol{\beta} = \mathbf{X}^{-1}\mathbf{Y}$ leads to $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, that is, ordinary least-squares will lead to a perfect fit, which is typically not good for generalization to unseen data). Also when $k > n$, then $\mathbf{X}^\top\mathbf{X}$ is not invertible and the normal equations admit a linear subspace of solutions. These behaviors of OLS in high dimension (k large) are often undesirable.

Several solutions exist to fix these issues. The most common is to regularize the least squares objective, either by adding an ℓ_1-penalty $\|\boldsymbol{\beta}\|_1$ to the empirical risk (leading to "Lasso" regression, see Chapter 8 of First Principles by Bach) or $\|\boldsymbol{\beta}\|_2^2$ (leading to ridge regression, as done in the following and also Chapter 7 of First Principles by Bach).

Definition 1.12.1 *(Ridge least-squares regression). For a regularization parameter $\lambda > 0$, we define the ridge least-squares estimator $\widehat{\boldsymbol{\beta}}_{ridge}$ as the minimizer of*

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2.$$

The ridge regression solution can be obtained in closed form,

$$\widehat{\boldsymbol{\beta}}_{ridge} = \left(\mathbf{X}^\top\mathbf{X} + n\lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{Y}.$$

As for the OLS estimator, we can analyze the statistical properties of this estimator under the linear model and fixed design assumptions. See Chapter 7 of First Principles by Bach for an analysis for random design and potentially infinite-dimensional features.

Proposition 1.12.2 *Recall that $\widehat{\Sigma} := \frac{1}{n}\mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{k\times k}$. Under the linear model assumption (and for the fixed design setting), the ridge least-squares estimator $\widehat{\boldsymbol{\beta}}_{ridge}$ has the following excess risk*

$$E[\mathcal{E}(\widehat{\boldsymbol{\beta}}_{ridge})] - \mathcal{E}_* = \lambda^2\boldsymbol{\beta}_*^\top(\widehat{\Sigma} + \lambda\mathbf{I})^{-2}\widehat{\Sigma}\boldsymbol{\beta}_* + \frac{\sigma^2}{n}\mathrm{tr}\left[\widehat{\Sigma}^2(\widehat{\Sigma} + \lambda\mathbf{I})^{-2}\right].$$

Proof. We use the risk decomposition of Proposition 1.9.1 into a bias term B and a variance term V. Since we have

$$E[\widehat{\boldsymbol{\beta}}_{\mathrm{ridge}}] = \frac{1}{n}(\widehat{\Sigma} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_* = (\widehat{\Sigma} + \lambda\mathbf{I})^{-1}\widehat{\Sigma}\boldsymbol{\beta}_* = \boldsymbol{\beta}_* - \lambda(\widehat{\Sigma} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}_*,$$

it follows,

$$B = \underbrace{\|E[\widehat{\boldsymbol{\beta}}_{\mathrm{ridge}}] - \boldsymbol{\beta}_*\|_{\widehat{\Sigma}}^2}_{\text{Bias}} = \lambda^2\boldsymbol{\beta}_*^\top(\widehat{\Sigma} + \lambda\mathbf{I})^{-2}\widehat{\Sigma}\boldsymbol{\beta}_*.$$

For the variance term, using the fact that $E[\epsilon\epsilon^\top] = \sigma^2$, we have

$$\begin{aligned}
V &= \underbrace{E\left[\|\widehat{\boldsymbol{\beta}}_{\text{ridge}} - E[\widehat{\boldsymbol{\beta}}_{\text{ridge}}]\|_{\widehat{\Sigma}}^2\right]}_{\text{Variance}} = E\left[\left\|\frac{1}{n}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\epsilon\right\|_{\widehat{\Sigma}}^2\right] \\
&= E\left[\frac{1}{n^2}\text{tr}\left(\epsilon^\top\mathbf{X}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\epsilon\right)\right] \\
&= E\left[\frac{1}{n^2}\text{tr}\left(\mathbf{X}^\top\epsilon\epsilon^\top\mathbf{X}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}\right)\right] = \frac{\sigma^2}{n}\text{tr}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}\right) \\
&= \frac{\sigma^2}{n}\text{tr}\left(\widehat{\Sigma}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}\right) = \frac{\sigma^2}{n}\text{tr}\left(\widehat{\Sigma}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}\widehat{\Sigma}\right) \\
&= \frac{\sigma^2}{n}\text{tr}\left[\widehat{\Sigma}^2(\widehat{\Sigma}+\lambda\mathbf{I})^{-2}\right] \qquad ((\widehat{\Sigma}+\lambda\mathbf{I})^{-1}\widehat{\Sigma} = \widehat{\Sigma}(\widehat{\Sigma}+\lambda\mathbf{I})^{-1}).
\end{aligned}$$

The proposition follows by summing the bias and variance terms. ■

We can make the following observations:

Remark 1.12.3

• *The result above is also a bias / variance decomposition with the bias term equal to* $B = \lambda^2\boldsymbol{\beta}_*^\top(\widehat{\Sigma}+\lambda\mathbf{I})^{-2}\widehat{\Sigma}\boldsymbol{\beta}_*$, *and the variance term equal to* $V = \frac{\sigma^2}{n}\text{tr}\left[\widehat{\Sigma}^2(\widehat{\Sigma}+\lambda\mathbf{I})^{-2}\right]$.

• *The bias term is increasing in* λ *and equal to zero for* $\lambda = 0$ *if* $\widehat{\Sigma}$ *is invertible, while when* λ *goes to infinity, the bias goes to* $\boldsymbol{\beta}_*^\top\widehat{\Sigma}\boldsymbol{\beta}_*$. *It is independent of* n *and plays the role of the approximation error in the risk decomposition.*

• *The variance term is decreasing in* λ, *and equal to* $\sigma^2 k/n$ *for* $\lambda = 0$ *and* $\widehat{\Sigma}$ *invertible, and converging to zero when* λ *goes to infinity. It depends on* n *and plays the role of the estimation error in the risk decomposition.*

• *The quantity* $\text{tr}[\widehat{\Sigma}^2(\widehat{\Sigma}+\lambda\mathbf{I})^{-2}]$ *is often called the "degrees of freedom", and is often considered as an implicit number of parameters. It can be expressed as where* $\sum_{j=1}^k \frac{\lambda_j^2}{(\lambda_j+\lambda)^2}$, *where* $(\lambda_j)_{j\in\{1,\dots,d\}}$ *are the eigenvalues of* $\widehat{\Sigma}$. *This quantity will be very important in the analysis of kernel methods in Chapter 7 of First Principles by Bach.*

• *Observe how this converges to the OLS estimator (when it is defined) as* $\lambda \to 0$.

• *In most cases,* $\lambda = 0$ *is not the optimal choice, that is biased estimation (with controlled bias) is preferable to unbiased estimation.*

Experiments

With the same polynomial regression set-up as in Bach book, with $k = 11$ (degree 10), we can plot the various quantities above as a function of λ. We can see the monotonicity of bias and variance with respect to λ as well as the presence of an optimal choice of λ. See Figure 1.5.

1.12.2 Choice of λ

Based on the expression for the risk, we can tune the regularization parameter λ to obtain a potentially better bound than with the OLS (which corresponds to $\lambda = 0$ and the excess risk $\sigma^2 k/n$).

Proposition 1.12.4 *(Choice of Regularization Parameter) With the choice* $\lambda^* = \frac{\sigma\sqrt{\text{tr}[\widehat{\Sigma}]}}{\|\boldsymbol{\beta}_*\|_2\sqrt{n}}$, *we have*

$$E[\mathcal{E}(\widehat{\boldsymbol{\beta}}_{ridge})] - \mathcal{E}_* \leq \frac{\sigma\sqrt{\text{tr}[\widehat{\Sigma}]}\|\boldsymbol{\beta}_*\|_2}{\sqrt{n}}.$$

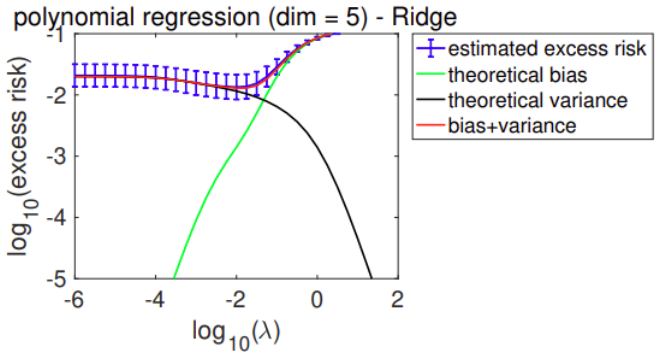


Figure 1.5: Bias-variance trade-offs for ridge regression.

Proof. We have, using the fact that the eigenvalues of $(\widehat{\Sigma}+\lambda\mathbf{I})^{-2}\lambda\widehat{\Sigma}$ are less than $1/2$ (which is a simple consequence of $(\mu+\lambda)^{-2}\mu\lambda \leq 1/2 \Leftrightarrow (\mu+\lambda)^2 \geq 2\mu\lambda$ for all eigenvalues μ of $\widehat{\Sigma}$):

$$B = \lambda^2\boldsymbol{\beta}_*^\top(\widehat{\Sigma}+\lambda\mathbf{I})^{-2}\widehat{\Sigma}\boldsymbol{\beta}_* = \lambda\boldsymbol{\beta}_*^\top(\widehat{\Sigma}+\lambda\mathbf{I})^{-2}\lambda\widehat{\Sigma}\boldsymbol{\beta}_* \leq \frac{\lambda}{2}\|\boldsymbol{\beta}_*\|_2^2.$$

Similarly, we have

$$V = \frac{\sigma^2}{n}\mathrm{tr}\left[\widehat{\Sigma}^2(\widehat{\Sigma}+\lambda\mathbf{I})^{-2}\right] = \frac{\sigma^2}{\lambda n}\mathrm{tr}\left[\widehat{\Sigma}\lambda\widehat{\Sigma}(\widehat{\Sigma}+\lambda\mathbf{I})^{-2}\right] \leq \frac{\sigma^2\mathrm{tr}[\widehat{\Sigma}]}{2\lambda n}.$$

Plugging in λ^* (which was chosen to minimize the upper bound on $B+V$) gives the result. ■

We can make the following observations:

Remark 1.12.5

• *Observe that if we write* $R = \max_{i\in\{1,\dots,n\}}\|\mathbf{X}_i\|_2$, *then we have*

$$\mathrm{tr}[\widehat{\Sigma}] = \sum_{j\geq 1}\widehat{\Sigma}_{jj} = \frac{1}{n}\sum_{i=1}^n\sum_{j\geq 1}x_{ij}^2 = \frac{1}{n}\sum_{i=1}^n\|\mathbf{X}_i\|_2^2 \leq R^2.$$

Thus in the excess risk bound, the dimension k *plays no role and it could even be infinite (given that* R *and* $\|\boldsymbol{\beta}_*\|_2$ *remain finite). This type of bounds are called* ***dimension-free*** *bounds. Notice that the number of parameters is not the only way to measure the generalization capabilities of a learning method*

• *Comparing this bound with that of the OLS estimator, we see that it converges slower to* 0 *as a function of* n *(from* n^{-1} *to* $n^{-1/2}$*) but it has a milder dependence on the noise (from* σ^2 *to* σ*). The presence of a "fast" rate in* $O(n^{-1})$ *with a potentially large constant, and of "slow" rate* $O(n^{-1/2})$ *with a smaller constant will appear several times. Notice that depending on* n *and the constants, the "fast" rate result is not always the best.*

• *The value of* λ^* *involves quantities which we typically do not know in practice (such as* σ *and* $\|\boldsymbol{\beta}_*\|_2$*). This is still useful to highlight the existence of some* λ *with good predictions (which can be found by cross-validation).*

• *Note here that the choice of* $\lambda^* = \frac{\sigma\sqrt{\mathrm{tr}[\widehat{\Sigma}]}}{\|\boldsymbol{\beta}_*\|_2\sqrt{n}}$ *is optimizing the* ***upper-bound*** $\frac{\lambda}{2}\|\boldsymbol{\beta}_*\|_2^2 + \frac{\sigma^2\mathrm{tr}[\widehat{\Sigma}]}{2\lambda n}$*, and is thus typically not optimal for the true expected risk.*

Choosing λ in practice. The regularization λ is an example of a **hyper-parameter**. This term refers broadly to any quantity that influences the behavior of a machine learning algorithm and that is left to choose

by the practitioner. While theory often offers guidelines and qualitative understanding on how to best choose the hyper-parameters, their precise numerical value depends on quantities which are often difficult to know or even guess. In practice, we typically resort to **validation and cross-validation**.

1.13 Subset Selection

Due to the time constraint, I have not enough time to well-organize the following section. In this section, I just copy-paste the content from the book "an introduction to statistical learning" by James, Witten, Hastie, Tibshirani, and the book "the elements of statistical learning" by Hastie, Tibshirani, Friedman. In the future, these contents need to be understood and typed in using latex.

1.14 Logistic Regression

Due to the time constraint, I have not enough time to well-organize the following section. In this section, I just copy-paste the content from the book "All of Statistics - A Concise Course in Statistical Inference" by Larry Wasserman, and the website "Logistic Regression" by Zhihu. In the future, these contents need to be understood and typed in using latex.

Notice that logistic regression is intrinsically regression during the computation procedure while its goal is for classification. On the other hand, support vector machine is completely for classification.

6.1 Subset Selection

In this section we consider some methods for selecting subsets of predictors. These include best subset and stepwise model selection procedures.

6.1.1 Best Subset Selection

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.

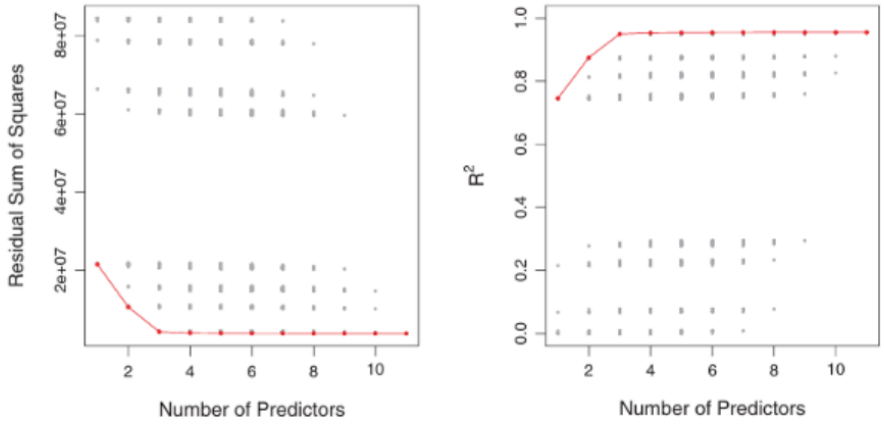


FIGURE 6.1. *For each possible model containing a subset of the ten predictors in the* `Credit` *data set, the RSS and* R^2 *are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and* R^2*. Though the data set contains only ten predictors, the* x*-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.*

Figure 1.6:

# Variables	Best subset	Forward stepwise
One	`rating`	`rating`
Two	`rating`, `income`	`rating`, `income`
Three	`rating`, `income`, `student`	`rating`, `income`, `student`
Four	`cards`, `income`	`rating`, `income`,
	`student`, `limit`	`student`, `limit`

TABLE 6.1. *The first four selected models for best subset selection and forward stepwise selection on the* `Credit` *data set. The first three models are identical but the fourth models differ.*

6.1.3 Choosing the Optimal Model

Best subset selection, forward selection, and backward selection result in the creation of a set of models, each of which contains a subset of the p predictors. In order to implement these methods, we need a way to determine which of these models is *best*. As we discussed in Section 6.1.1, the model containing all of the predictors will always have the smallest RSS and the largest R^2, since these quantities are related to the training error. Instead, we wish to choose a model with a low test error. As is evident here, and as we show in Chapter 2, the training error can be a poor estimate of the test error. Therefore, RSS and R^2 are not suitable for selecting the best model among a collection of models with different numbers of predictors.

In order to select the best model with respect to test error, we need to estimate this test error. There are two common approaches:

1. We can indirectly estimate test error by making an *adjustment* to the training error to account for the bias due to overfitting.

2. We can *directly* estimate the test error, using either a validation set approach or a cross-validation approach, as discussed in Chapter 5.

We consider both of these approaches below.

Figure 1.7:

C_p, AIC, BIC, and Adjusted R^2

We show in Chapter 2 that the training set MSE is generally an underestimate of the test MSE. (Recall that MSE = RSS/n.) This is because when we fit a model to the training data using least squares, we specifically estimate the regression coefficients such that the training RSS (but not the test RSS) is as small as possible. In particular, the training error will decrease as more variables are included in the model, but the test error may not. Therefore, training set RSS and training set R^2 cannot be used to select from among a set of models with different numbers of variables.

However, a number of techniques for *adjusting* the training error for the model size are available. These approaches can be used to select among a set of models with different numbers of variables. We now consider four such approaches: C_p, *Akaike information criterion* (AIC), *Bayesian information criterion* (BIC), and *adjusted* R^2. Figure 6.2 displays C_p, BIC, and adjusted R^2 for the best model of each size produced by best subset selection on the `Credit` data set.

C_p

Akaike information criterion

Bayesian information criterion

adjusted R^2

For a fitted least squares model containing d predictors, the C_p estimate of test MSE is computed using the equation

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right), \tag{6.2}$$

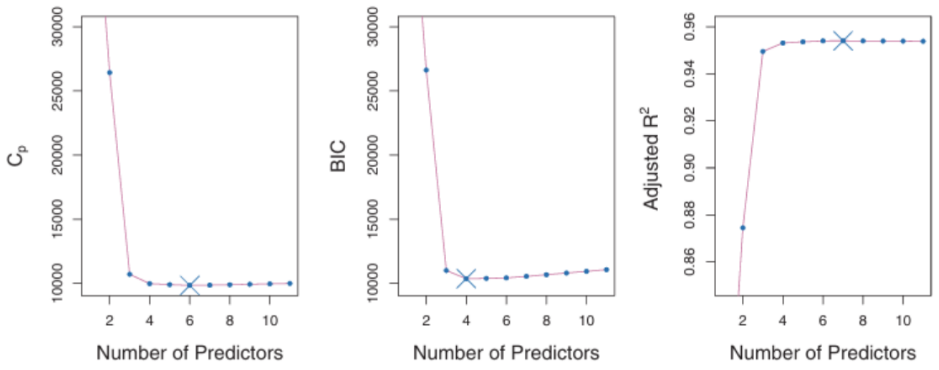


FIGURE 6.2. *C_p, BIC, and adjusted R^2 are shown for the best models of each size for the `Credit` data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.*

Figure 1.8:

where $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement in (6.1).[3] Essentially, the C_p statistic adds a penalty of $2d\hat{\sigma}^2$ to the training RSS in order to adjust for the fact that the training error tends to underestimate the test error. Clearly, the penalty increases as the number of predictors in the model increases; this is intended to adjust for the corresponding decrease in training RSS. Though it is beyond the scope of this book, one can show that if $\hat{\sigma}^2$ is an unbiased estimate of σ^2 in (6.2), then C_p is an unbiased estimate of test MSE. As a consequence, the C_p statistic tends to take on a small value for models with a low test error, so when determining which of a set of models is best, we choose the model with the lowest C_p value. In Figure 6.2, C_p selects the six-variable model containing the predictors `income`, `limit`, `rating`, `cards`, `age` and `student`.

The AIC criterion is defined for a large class of models fit by maximum likelihood. In the case of the model (6.1) with Gaussian errors, maximum likelihood and least squares are the same thing. In this case AIC is given by

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + 2d\hat{\sigma}^2\right),$$

where, for simplicity, we have omitted an additive constant. Hence for least squares models, C_p and AIC are proportional to each other, and so only C_p is displayed in Figure 6.2.

BIC is derived from a Bayesian point of view, but ends up looking similar to C_p (and AIC) as well. For the least squares model with d predictors, the BIC is, up to irrelevant constants, given by

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n)d\hat{\sigma}^2\right). \tag{6.3}$$

Like C_p, the BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value. Notice that BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $\log(n)d\hat{\sigma}^2$ term, where n is the number of observations. 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. In Figure 6.2, we see that this is indeed the case for the `Credit` data set; BIC chooses a model that contains only the four predictors `income`, `limit`, `cards`, and `student`. In this case the curves are very flat and so there does not appear to be much difference in accuracy between the four-variable and six-variable models.

Figure 1.9:

The adjusted R^2 statistic is another popular approach for selecting among a set of models that contain different numbers of variables. Recall from Chapter 3 that the usual R^2 is defined as $1 - \text{RSS}/\text{TSS}$, where $\text{TSS} = \sum(y_i - \overline{y})^2$ is the *total sum of squares* for the response. Since RSS always decreases as more variables are added to the model, the R^2 always increases as more variables are added. For a least squares model with d variables, the adjusted R^2 statistic is calculated as

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}. \tag{6.4}$$

Unlike C_p, AIC, and BIC, for which a *small* value indicates a model with a low test error, a *large* value of adjusted R^2 indicates a model with a small test error. Maximizing the adjusted R^2 is equivalent to minimizing $\frac{\text{RSS}}{n-d-1}$. While RSS always decreases as the number of variables in the model increases, $\frac{\text{RSS}}{n-d-1}$ may increase or decrease, due to the presence of d in the denominator.

The intuition behind the adjusted R^2 is that once all of the correct variables have been included in the model, adding additional *noise* variables

will lead to only a very small decrease in RSS. Since adding noise variables leads to an increase in d, such variables will lead to an increase in $\frac{\text{RSS}}{n-d-1}$, and consequently a decrease in the adjusted R^2. Therefore, in theory, the model with the largest adjusted R^2 will have only correct variables and no noise variables. Unlike the R^2 statistic, the adjusted R^2 statistic *pays a price* for the inclusion of unnecessary variables in the model. Figure 6.2 displays the adjusted R^2 for the `Credit` data set. Using this statistic results in the selection of a model that contains seven variables, adding `gender` to the model selected by C_p and AIC.

C_p, AIC, and BIC all have rigorous theoretical justifications that are beyond the scope of this book. These justifications rely on asymptotic arguments (scenarios where the sample size n is very large). Despite its popularity, and even though it is quite intuitive, the adjusted R^2 is not as well motivated in statistical theory as AIC, BIC, and C_p. All of these measures are simple to use and compute. Here we have presented the formulas for AIC, BIC, and C_p in the case of a linear model fit using least squares; however, these quantities can also be defined for more general types of models.

Figure 1.10:

Validation and Cross-Validation

As an alternative to the approaches just discussed, we can directly estimate the test error using the validation set and cross-validation methods discussed in Chapter 5. We can compute the validation set error or the cross-validation error for each model under consideration, and then select the model for which the resulting estimated test error is smallest. This procedure has an advantage relative to AIC, BIC, C_p, and adjusted R^2, in that it provides a direct estimate of the test error, and makes fewer assumptions about the true underlying model. It can also be used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom (e.g. the number of predictors in the model) or hard to estimate the error variance σ^2.

In the past, performing cross-validation was computationally prohibitive for many problems with large p and/or large n, and so AIC, BIC, C_p, and adjusted R^2 were more attractive approaches for choosing among a set of models. However, nowadays with fast computers, the computations required to perform cross-validation are hardly ever an issue. Thus, cross-validation is a very attractive approach for selecting from among a number of models under consideration.

Figure 6.3 displays, as a function of d, the BIC, validation set errors, and cross-validation errors on the `Credit` data, for the best d-variable model. The validation errors were calculated by randomly selecting three-quarters of the observations as the training set, and the remainder as the validation set. The cross-validation errors were computed using $k = 10$ folds. In this case, the validation and cross-validation methods both result in a six-variable model. However, all three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.

Figure 1.11:

In fact, the estimated test error curves displayed in the center and right-hand panels of Figure 6.3 are quite flat. While a three-variable model clearly has lower estimated test error than a two-variable model, the estimated test errors of the 3- to 11-variable models are quite similar. Furthermore, if we repeated the validation set approach using a different split of the data into a training set and a validation set, or if we repeated cross-validation using a different set of cross-validation folds, then the precise model with the lowest estimated test error would surely change. In this setting, we can select a model using the *one-standard-error rule*. We first calculate the standard error of the estimated test MSE for each model size, and then select the smallest model for which the estimated test error is within one

one-standard-error rule

standard error of the lowest point on the curve. The rationale here is that if a set of models appear to be more or less equally good, then we might as well choose the simplest model—that is, the model with the smallest number of predictors. In this case, applying the one-standard-error rule to the validation set or cross-validation approach leads to selection of the three-variable model.

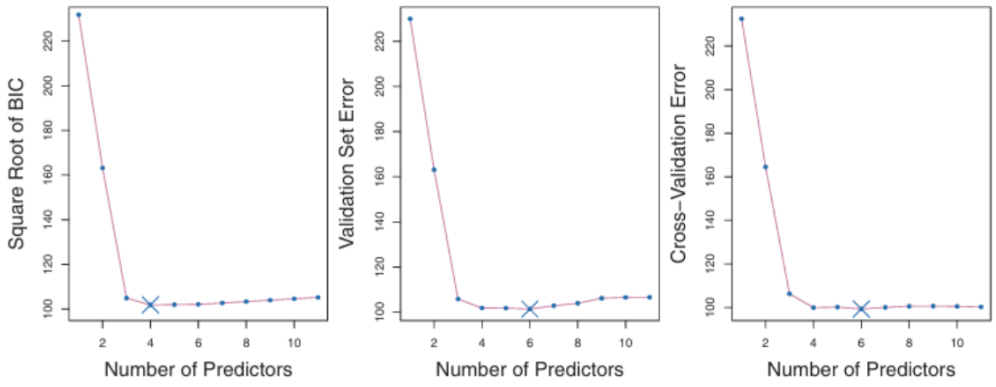


FIGURE 6.3. *For the* `Credit` *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11. *The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC.* Center: *Validation set errors.* Right: *Cross-validation errors.*

Figure 1.12:

More reads

AIC, BIC, Logistic. see Hastie An Introduction to statistical learning. and the elements of Stat. Learning.

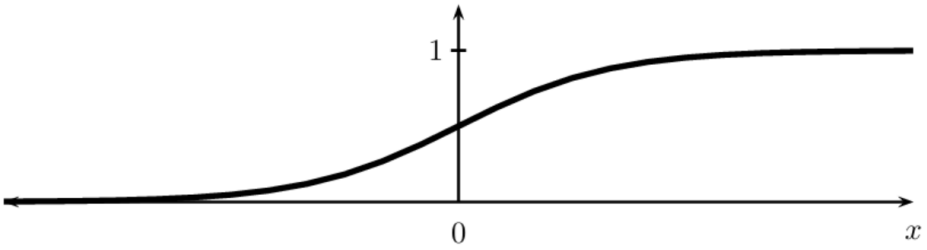


FIGURE 13.3. The logistic function $p = e^x/(1 + e^x)$.

13.7 Logistic Regression

So far we have assumed that Y_i is real valued. **Logistic regression** is a parametric method for regression when $Y_i \in \{0, 1\}$ is binary. For a k-dimensional covariate X, the model is

$$p_i \equiv p_i(\beta) \equiv \mathbb{P}(Y_i = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1}^k \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^k \beta_j x_{ij}}} \tag{13.32}$$

or, equivalently,

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} \tag{13.33}$$

where

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right). \tag{13.34}$$

Figure 1.13:

The name "logistic regression" comes from the fact that $e^x/(1+e^x)$ is called the logistic function. A plot of the logistic for a one-dimensional covariate is shown in Figure 13.3.

Because the Y_i's are binary, the data are Bernoulli:

$$Y_i|X_i = x_i \sim \text{Bernoulli}(p_i).$$

Hence the (conditional) likelihood function is

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} p_i(\beta)^{Y_i}(1-p_i(\beta))^{1-Y_i}. \tag{13.35}$$

The MLE $\widehat{\beta}$ has to be obtained by maximizing $\mathcal{L}(\beta)$ numerically. There is a fast numerical algorithm called reweighted least squares. The steps are as follows:

Reweighted Least Squares Algorithm

Choose starting values $\widehat{\beta}^0 = (\widehat{\beta}_0^0, \ldots, \widehat{\beta}_k^0)$ and compute p_i^0 using equation (13.32), for $i = 1, \ldots, n$. Set $s = 0$ and iterate the following steps until convergence.

1. Set
$$Z_i = \text{logit}(p_i^s) + \frac{Y_i - p_i^s}{p_i^s(1-p_i^s)}, \quad i = 1, \ldots, n.$$

2. Let W be a diagonal matrix with (i,i) element equal to $p_i^s(1-p_i^s)$.

3. Set
$$\widehat{\beta}^s = (X^T W X)^{-1} X^T W Y.$$
This corresponds to doing a (weighted) linear regression of Z on Y.

4. Set $s = s+1$ and go back to the first step.

Figure 1.14:

用人话讲明白逻辑回归Logistic regression

化简可得

+ 关注她

2582 人赞同了该文章

文章目录

今天梳理一下逻辑回归，这个算法由于简单、实用、高效，在业界应用十分广泛。注意咯，这里的“逻辑”是音译“逻辑斯蒂（logistic）”的缩写，并不是说这个算法具有怎样的逻辑性。

前面说过，机器学习算法中的监督式学习可以分为2大类：

- **分类模型**：目标变量是分类变量（离散值）；
- **回归模型**：目标变量是连续性数值变量。

逻辑回归通常用于解决分类问题，例如，业界经常用它来预测：客户是否会购买某个商品，借款人是否会违约等等。

实际上，**“分类”是应用逻辑回归的目的和结果，但中间过程依旧是“回归”**。

为什么这么说?

因为通过逻辑回归模型，我们得到的计算结果是0-1之间的连续数字，可以把它称为“**可能性**”（概率）。对于上述问题，就是：客户购买某个商品的可能性，借款人违约的可能性。

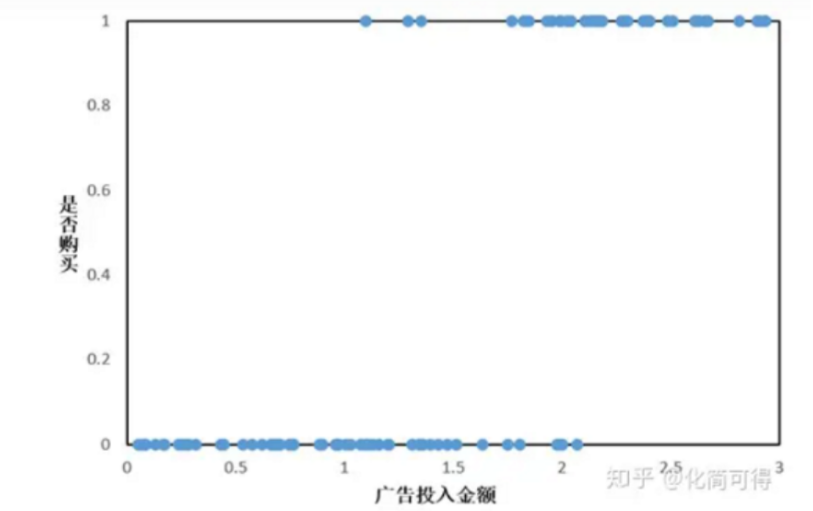
然后，给这个可能性加一个阈值，就成了分类。例如，算出贷款违约的可能性>0.5，将借款人预判为坏客户。

Figure 1.15:

1.从线性回归说起

考虑最简单的情况，即只有一个自变量的情况。比方说广告投入金额x和销售量y的关系，散点图✦如下，这种情况适用一元线性回归✦。

但在许多实际问题中，因变量✦y是分类型，只取0、1两个值，和x的关系不是上面那样。假设我们有这样一组数据：给不同的用户投放不同金额的广告，记录他们购买广告商品的行为，1代表购买，0代表未购买。



假如此时依旧考虑线性回归模型✦，得到如下拟合曲线：

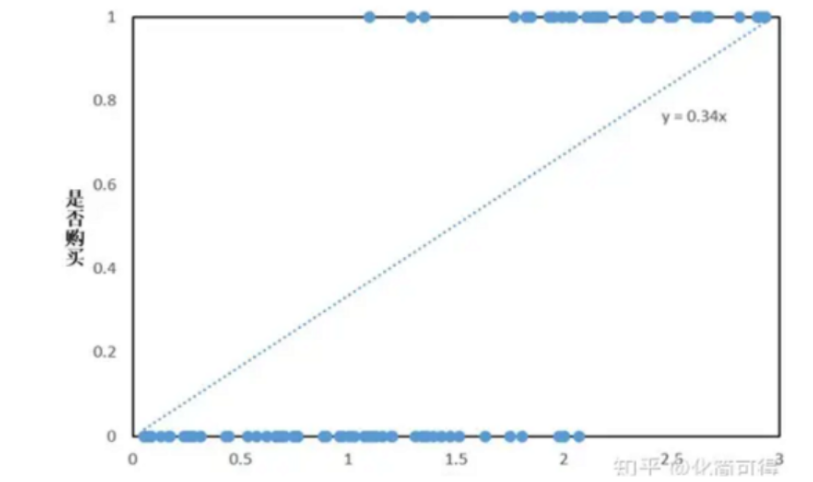


Figure 1.16:

线性回归拟合的曲线，看起来和散点毫无关系，似乎没有意义。但我们可以在计算出 $\hat{y}$ 的结果后，加一个限制，即 $\hat{y} > 0.5$ ，就认为其属于1这一类，购买了商品，否则认为其不会购买，即：

$$\hat{y} = \begin{cases} 1, f(x) > 0.5 \\ 0, f(x) \leq 0.5 \end{cases}$$

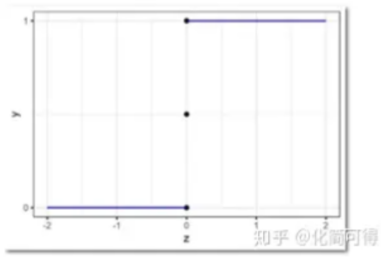
由于拟合方程为 $\hat{y} = 0.34 * x$ ，那么上面的限制就等价于：

$$\hat{y} = \begin{cases} 1, x > 1.47 \\ 0, x \leq 1.47 \end{cases}$$

这种形式，非常像单位阶跃函数：

$$y = \begin{cases} 0, z < 0 \\ 0.5, z = 0 \\ 1, z > 0 \end{cases}$$

图像如下：



我们发现，把阶跃函数向右平移一下，就可以比较好地拟合上面的散点图呀！但是阶跃函数有个问题，它不是连续函数。

理想的情况，是像线性回归的函数一样，**X和Y之间的关系，是用一个单调可导的函数来描述的。**

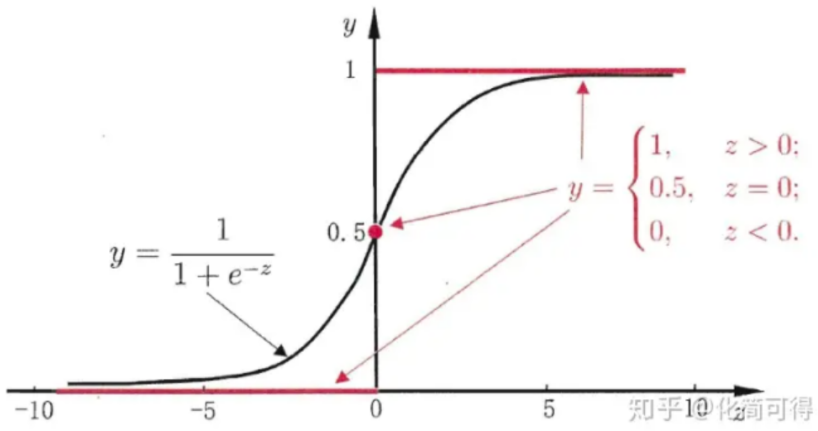
Figure 1.17:

2.sigmond函数

实际上，逻辑回归算法*的拟合函数，叫做sigmond函数：

$$f(z) = \frac{1}{1+e^{-z}}$$

函数图像如下（百度图片搜到的图）：



sigmoid函数是一个s形曲线，就像是阶跃函数的温和版，阶跃函数在0和1之间是突然的起跳，而sigmoid有个平滑的过渡。

从图形上看，sigmoid曲线就像是被掰弯捋平后的线性回归直线，**将取值范围(−∞,+∞)映射到(0,1)之间，更适宜表示预测的概率，即事件发生的“可能性”** 。

3.推广至多元场景

在用人话讲明白梯度下降Gradient Descent一文中，我们讲了多元线性回归方程的一般形式为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

可以简写为矩阵*形式： $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$

其中，

Figure 1.18:

$$\boldsymbol{Y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X}=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

将特征加权求和Xβ（后面不对矩阵向量⁺加粗了，大家应该都能理解）代入sigmond函数中的z，得到 $\frac{1}{1+e^{-X\beta}}$ ，令其为预测为正例的概率P(Y=1)，那么逻辑回归的形式就有了：

$$P(Y=1)=\frac{1}{1+e^{-X\beta}}$$

到目前为止，逻辑函数⁺的构造算是完成了。找到了合适的函数，下面就是求函数中的未知参数向量β了。求解之前，我们需要先理解一个概念——**似然性。**

4.似然函数

我们常常用**概率(Probability)** 来描述一个事件发生的可能性。

而**似然性(Likelihood)** 正好反过来，意思是一个事件实际已经发生了，反推在什么参数条件下，这个事件发生的概率最大。

用数学公式来表达上述意思，就是:

- 已知参数 β 前提下，预测某事件 x 发生的条件概率为 $P(x|\beta)$；
- 已知某个已发生的事件 x，未知参数 β 的似然函数为 $\mathcal{L}(\beta|x)$；
- 上面两个值相等，即: $\mathcal{L}(\beta|x)=P(x|\beta)$ 。

一个参数 β 对应一个似然函数的值，当 β 发生变化， $\mathcal{L}(\beta|x)$ 也会随之变化。当我们在取得某个参数的时候，似然函数的值到达了最大值，说明在这个参数下最有可能发生x事件，即这个参数最合理。

因此，最优β，就是使当前观察到的数据出现的可能性最大的β。

5.最大似然估计⁺

在二分类问题中，y只取0或1，可以组合起来表示y的概率:

$$P(y)=P(y=1)^{y}P(y=0)^{1-y}$$

我们可以把y=1代入上式验证下:

Figure 1.19:

- 左边是P(y=1)；
- 右边是 $P(y=1)^1P(y=0)^0$ ，也为P(y=1)。

上面的式子，更严谨的写法需要加上特征x和参数β：

$$P(y|x,\beta)=P(y=1|x,\beta)^y[1-P(y=1|x,\beta)]^{1-y}$$

前面说了， $\frac{1}{1+e^{-x\beta}}$ 表示的就是P(y=1)，代入上式：

$$P(y|x,\beta)=\left(\frac{1}{1+e^{-x\beta}}\right)^y\left(1-\frac{1}{1+e^{-x\beta}}\right)^{1-y}$$

根据上一小节说的最优β的定义，也就是最大化我们见到的样本数据*的概率，即求下式的最大值。

$$\mathcal{L}(\beta)=\prod_{i=1}^{n}P(y_i|x_i,\beta)=\prod_{i=1}^{n}\left(\frac{1}{1+e^{-x_i\beta}}\right)^{y_i}\left(1-\frac{1}{1+e^{-x_i\beta}}\right)^{1-y_i}$$

这个式子怎么来的呢？

其实很简单。

前面我们说了， $\mathcal{L}(\beta|x)=P(x|\beta)$ ，对于某个观测值yi，似然函数的值 $\mathcal{L}(\beta|y_i)$ ，就等于条件概率*的值 $P(y_i|\beta)$ 。

另外我们知道，如果事件A与事件B相互独立，那么两者同时发生的概率为P(A)*P(B)。那么我们观测到的y1,y2.....yn，他们同时发生的概率就是 $\prod_{i=1}^{n}P(y_i|\beta)$ 。

因为一系列的xi和yi都是我们实际观测到的数据，式子中未知的只有β。因此，现在问题就变成了**求β在取什么值的时候，L(β)能达到最大值。**

L(β)是所有观测到的y发生概率的乘积，这种情况求最大值比较麻烦，一般我们会先取对数，将乘积转化成加法。

取对数后，转化成下式：

$$log\mathcal{L}(\beta)=\sum_{i=1}^{n}\left([y_i\cdot log(\frac{1}{1+e^{-x_i\beta}})]+[(1-y_i)\cdot log(1-\frac{1}{1+e^{-x_i\beta}})]\right)$$

接下来想办法求上式的最大值就可以了，求解前，我们要提一下逻辑回归的损失函数。

6.损失函数

在机器学习领域，总是避免不了谈论损失函数这一概念。**损失函数是用于衡量预测值与实际值的偏离程度，即模型预测的错误程度。**也就是说，这个值越小，认为模型效果越好，举个极端例子，如果预测完全精确，则损失函数值为0。

Figure 1.20:

在线性回归一文中，我们用到的损失函数是残差平方和SSE：

$$Q = \sum_1^n (y_i - \hat{y}_i)^2 = \sum_1^n (y_i - x_i\beta)^2$$

这是个凸函数，有全局最优解。

如果逻辑回归也用平方损失，那么就是：

$$Q = \sum_1^n \left(y_i - \frac{1}{1+e^{-x_i\beta}}\right)^2$$

很遗憾，这个不是凸函数，不易优化，容易陷入局部最小值，所以逻辑函数用的是别的形式的函数作为损失函数，叫**对数损失函数**（log loss function）。

这个对数损失，就是上一小节的**似然函数取对数后，再取相反数**哟：

$$J(\beta) = -log\mathcal{L}(\beta) = -\sum_{i=1}^n [y_i logP(y_i) + (1-y_i)log(1-P(y_i))]$$

这个对数损失函数好理解吗？我还是举个具体例子吧。

用文章开头那个例子，假设我们有一组样本，建立了一个逻辑回归模型P(y=1)=f(x)，其中一个样本A是这样的：

之前我们在用人话讲明白梯度下降中解释过梯度下降算法，下面我们就用梯度下降法求损失函数的最小值（也可以用梯度上升算法求似然函数的最大值，这两是等价的）。

7.梯度下降法求解

要开始头疼的公式推导部分了，不要害怕哦，我们还是从最简单的地方开始，非常容易看懂。

首先看，对于sigmoid函数 $f(x) = \frac{1}{1+e^{-x}}$ ，f'(x)等于多少？

如果你还记得导数表中这2个公式，那就好办了（不记得也没关系，这就给你列出来）：

$$(\frac{1}{x})' = -\frac{1}{x^2}$$

$$(e^x)' = e^x$$

根据上两个公式，推导：

$$f'(x) = (\frac{1}{1+e^{-x}})' = -\frac{(e^{-x})'}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2}$$

Figure 1.21:

到这还不算完哦，我们发现 $1-f(x)=\frac{e^{-x}}{1+e^{-x}}$ ，而f'(x)正好可以拆分为 $\frac{e^{-x}}{1+e^{-x}}\cdot\frac{1}{1+e^{-x}}$ ，也就是说：

$$f'(x)=f(x)\cdot(1-f(x))$$

当然，现在我们的x是已知的，未知的是β，所以后面是对β求导，记：

$$\frac{1}{1+e^{-x_i\beta}}=f(x_i\beta)$$

把它代入前面我们得到逻辑回归的损失函数✦：

$$J(\beta)=-\sum_{i=1}^{n}\left[y_i log(f(x_i\beta))+(1-y_i)log(1-f(x_i\beta))\right]=-\sum_{i=1}^{n}g(\beta_i,x_i,y_i)$$

简便起见，先撇开求和号看g(β,x,y)。不过这个g(β,x,y)里面也挺复杂的，我们再把里面的 $f(x_i\beta)$ 挑出来，单独先看它对β向量中的某个βj求偏导是什么样。

根据上面的求导公式✦，有：

$$\frac{\partial f(x_i\beta)}{\partial\beta_j}=f(x_i\beta)\cdot(1-f(x_i\beta))\cdot x_{ij}$$

注意咯，这个xi实际上指的是第i个样本的特征向量✦，即 $(1,x_{i1},\ldots,x_{ip})$ ，其中只有xij会和βj相乘，因此求导✦后整个xi只剩xij了。

理解了前面说的，下面的化简就轻而易举（知乎不能正常显示这个公式，只好写完转成图片粘过来）：

$$\begin{aligned}\frac{\partial g(\beta_i,x_i,y_i)}{\partial\beta_j}&=y_i\frac{1}{f(x_i\beta)}\frac{\partial f(x_i\beta)}{\partial\beta_j}-(1-y_i)\frac{1}{1-f(x_i\beta)}\frac{\partial f(x_i\beta)}{\partial\beta_j}\\&=\left(\frac{y_i}{f(x_i\beta)}-\frac{1-y_i}{1-f(x_i\beta)}\right)\cdot\frac{\partial f(x_i\beta)}{\partial\beta_j}\\&=\left(\frac{y_i}{f(x_i\beta)}-\frac{1-y_i}{1-f(x_i\beta)}\right)\cdot f(x_i\beta)\cdot(1-f(x_i\beta))\cdot x_{ij}\\&=[y_i(1-f(x_i\beta))-(1-y_i)f(x_i\beta)]\cdot x_{ij}\\&=(y_i-f(x_i\beta))\cdot x_{ij}\end{aligned}$$

Figure 1.22:

加上求和号：

$$\frac{\partial J(\beta)}{\partial \beta_j} = -\sum_{i=1}^{n}\left(y_i - f(x_i\beta)\right)\cdot x_{ij} = \sum_{i=1}^{n}\left(\frac{1}{1+e^{-x_i\beta}} - y_i\right)\cdot x_{ij}$$

有了偏导✦，也就有了梯度G，即偏导函数组成的向量。

梯度下降算法过程：

1. 初始化β向量的值，即 Θ_0 ，将其代入G得到当前位置的梯度；

2. 用步长α乘以当前梯度，得到从当前位置下降的距离；

3. 更新 Θ_1 ，其更新表达式✦为 $\Theta_1 = \Theta_0 - \alpha G$ ；

4. 重复以上步骤，直到更新到某个 Θ_k ，达到停止条件，这个 Θ_k 就是我们求解的参数向量。

8.结尾

文章写到这里就结束了，其实逻辑回归✦还有很多值得深入学习和讨论的，但是“讲人话”系列的定位就是个入门，我自己目前的水平也有限，所以本篇不再往后写了。

主要一篇文章写起来蛮累……知乎的编辑器真是很难用，每次都要点插入公式，而且好多语法还不能正常显示>.<

最后，欢迎各位一起学习讨论~

Figure 1.23: