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 26 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 $\mathrm{RSS} = \sum_{i=1}^n \widehat{\epsilon}_i^2$.

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

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

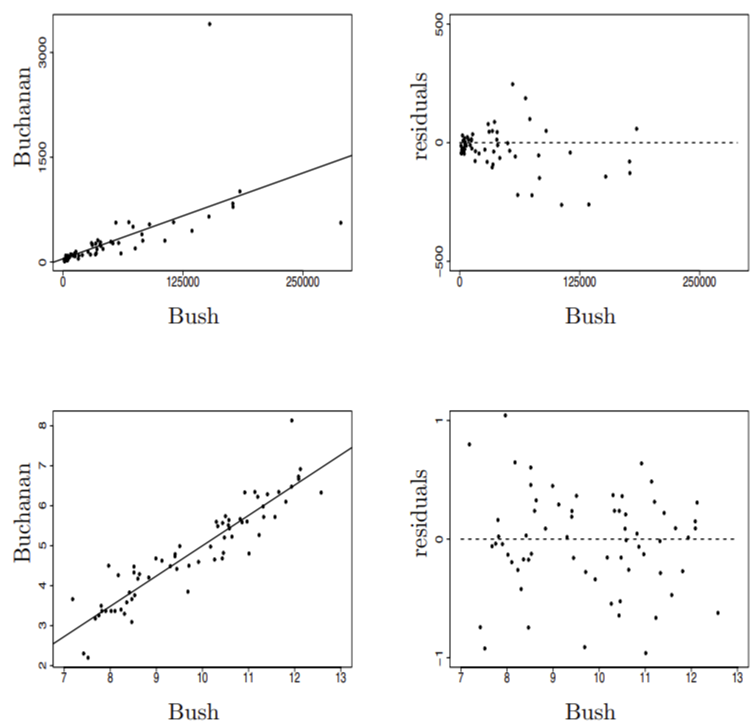

FIGURE 13.2. Voting Data for Election 2000. See example 13.6.

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*

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

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

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

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

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

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

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

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):

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

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

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

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 pair-wise 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

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

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

$$\begin{aligned}
E(Y_i-\overline{Y})^2 &= Var(Y_i-\overline{Y})+\left[E(Y_i-\overline{Y})\right]^2 \\
&= Var\left[\left(1-\frac{1}{n}\right)Y_i-\frac{1}{n}\sum_{j=1,j\neq i}^{n}Y_j\right]+\left[\beta_0+\beta_1X_i-\beta_0-\beta_1\overline{X}\right]^2 \\
&= \left(1-\frac{1}{n}\right)^2\sigma^2+\frac{(n-1)\sigma^2}{n^2}+\beta_1^2\left(X_i-\overline{X}\right)^2 \\
&= \left(1-\frac{1}{n}\right)\sigma^2+\beta_1^2\left(X_i-\overline{X}\right)^2.
\end{aligned}$$

For the second term, we use the following result,

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

Then

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

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

1.7 Prediction

Suppose we have estimated a regression model $\widehat{f}(x)=\widehat{\beta}_0+\widehat{\beta}_1x$ 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

$$\widehat{Y}_*=\widehat{\beta}_0+\widehat{\beta}_1x_*. \tag{1.15}$$

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

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

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

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

Thus the pivot quantity can be simplified,

$$\frac{\frac{Y-\widehat{\beta}_0-\widehat{\beta}_1 X}{\sqrt{Var(Y-\widehat{\beta}_0-\widehat{\beta}_1 X)}}}{\sqrt{\frac{1}{n-2}\frac{\text{RSS}}{\sigma^2}}}=\frac{\frac{Y-\widehat{\beta}_0-\widehat{\beta}_1 X}{\sigma\sqrt{1+\frac{\sum_{i=1}^n (X_i-X)^2}{n\sum_{i=1}^n (X_i-\overline{X})^2}}}}{\frac{\widehat{\sigma}}{\sigma}}=\frac{Y-\widehat{\beta}_0-\widehat{\beta}_1 X}{\widehat{\sigma}\sqrt{1+\frac{\sum_{i=1}^n (X_i-X)^2}{n\sum_{i=1}^n (X_i-\overline{X})^2}}}.$$

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

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

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

$$\mathbf{Y}=\begin{pmatrix} Y_1\\ \vdots\\ Y_n\end{pmatrix}\in\mathbb{R}^{n\times 1},$$

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

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

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,

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

■

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)* $\text{RSS}/\sigma^2 \sim \chi^2(n-q)$, *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,

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

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

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

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

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

Therefore,

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

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,

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

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 $\widehat{\beta}_1, \dots, \widehat{\beta}_p$ 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,*

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

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 $X_1, \dots, X_p$. If there is no linear relation between them, then all the β_j ($j = 1, \dots, p$) 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 $X_1, \dots, X_p$.

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

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

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 $j = 1, \dots, p$.

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

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

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.

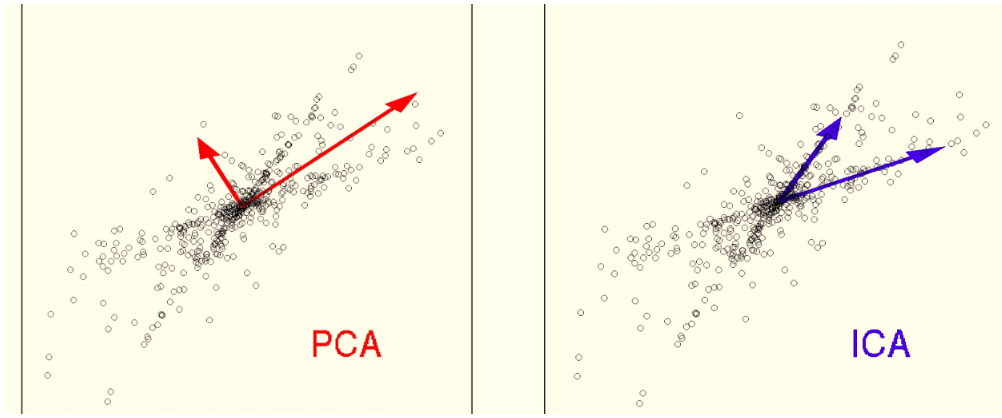

Figure 1.2: PCA vs. ICA.

• 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,

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

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

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

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

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

First we compute λ_0 and λ_1,

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

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

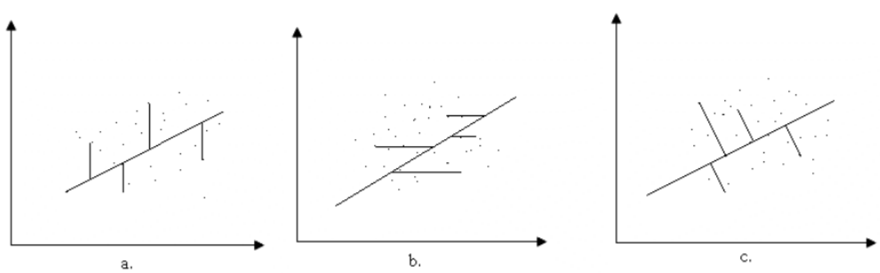

Figure 1.3: The OLS regression (a) and (b) and the OR (c).

As illustrated in Figure 1.3(a), the ordinary least square (OLS) estimate of Y on X will minimize the squared vertical distance $\sum_{i=1}^{n}\left(y_i - \alpha - \beta x_i\right)^2$ 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 $\sum_{i=1}^{n}\left(x_i - \alpha - \beta y_i\right)^2$ (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 $D_{XX} = \sum x_i^2, D_{XY} = \sum x_i y_i, D_{YY} = \sum y_i^2$,

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

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.

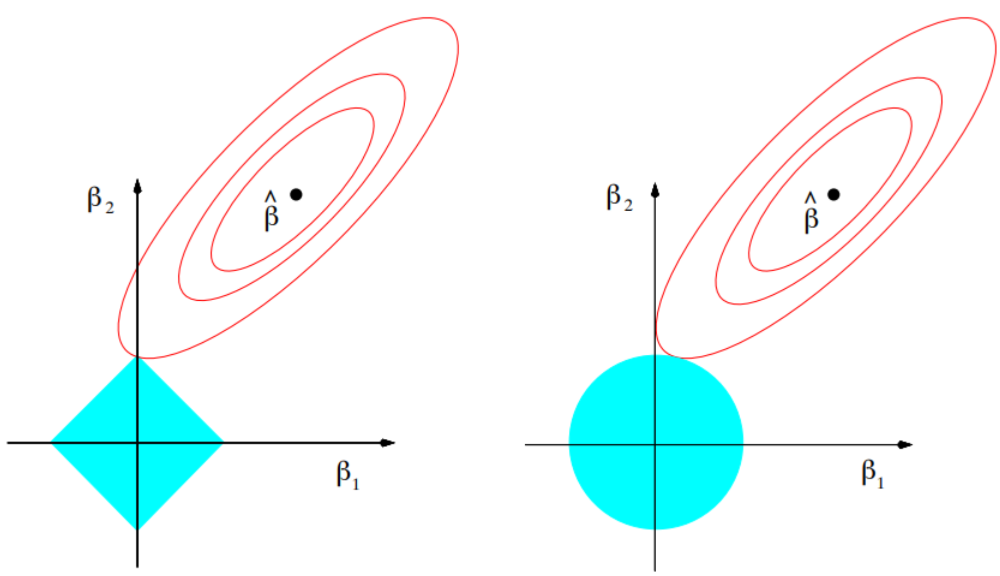

Figure 1.4: The "classical" illustration comparing lasso and ridge constraints. See Chap. 3.4 of Hastie et al. (2009).

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

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

it follows,

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

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

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

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* $V = \frac{\sigma^2}{n}\operatorname{tr}\left[\widehat{\Sigma}^2(\widehat{\Sigma}+\lambda\mathbf{I})^{-2}\right]$.
- *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* σ^2k/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* $\operatorname{tr}[\widehat{\Sigma}^2(\widehat{\Sigma}+\lambda\mathbf{I})^{-2}]$ *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 σ^2k/n).

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

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

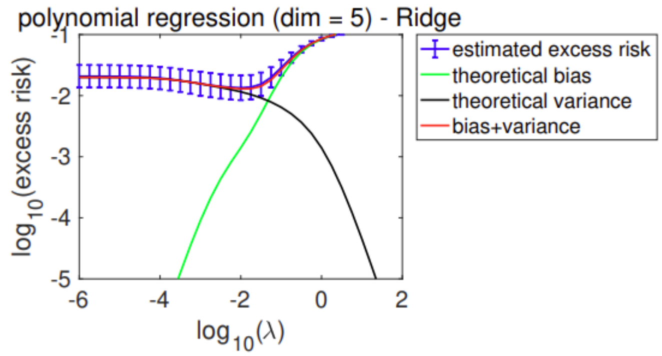

Figure 1.5: Bias-variance trade-offs for ridge regression.

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* $R = \max_{i\in\{1,\ldots,n\}}\|\mathbf{X}_i\|_2$*, 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**.