Notes on Statistical Learning

Mohammad Safdari

Contents

Chapter 1

Introduction to Statistical Learning

1.1 Populations, Samples, and Learning methods

Suppose the probability measure $\mathbb{P}$ describes the distribution of some quantities or properties among a **population**. Let Y, X be two of these quantities or properties. Mathematically, Y, X are random variables. They can also be vector-valued. There are many situations in which we are interested in predicting the value of Y based on the observed value of X, or understanding the relationship between Y, X. Suppose

$$Y = f(X) + \epsilon,$$

where ϵ is a random noise with $\mathbb{E}[\epsilon] = 0$. We usually assume that Y is scalar-valued, but $X = (X_1, \ldots, X_p)$ can be vector-valued. In this setting, the components of X are called **(input) variables**, or **predictors**, or **features**. And Y is called the **output variable**, or **response**.

To understand the relationship between Y, X, and to perform predictions, we need to estimate the function f. In order to estimate f, we need a **sample** from the population, which means we need a set of observations of the variables of interest X, Y. Suppose we have a data set consisting of n observations

$$(x_1, y_1), \ldots, (x_n, y_n)$$

for X, Y. Then we can employ this sample, which is also called the **training data**, to estimate f using a **learning method**. Broadly speaking, there are two types of learning methods for estimation of f. In **parametric methods** we first assume that f has a given form with some unknown parameters; then we use the data to estimate those parameters. For example, we may assume that f is a linear function; then we can try to estimate the coefficients of that linear function using the data.

In contrast, in **non-parametric methods** we do not assume a predetermined form for f.

The above process of estimating f is known as **supervised learning**, since we have both the response Y and the predictor X. In supervised learning, when the response Y is a **quantitative** variable (i.e. it takes numerical values), we are dealing with a **regression** problem. And when the response Y is a **qualitative** or **categorical** variable (i.e. it takes finitely many discrete values, in other words, its values belongs to one of finitely many **classes** or **categories**), we are dealing with a **classification** problem.

Sometimes there is no response variable at all, and we have only observed the input variables. In these settings we may for example be interested in detecting patterns in the data. This is known as **unsupervised learning**.

1.2 Measuring the Error

1.2.1 Reducible and Irreducible Errors

Suppose that we have estimated the following form for f:

$$\hat{Y} = \hat{f}(X).$$

Then the difference between $Y, \hat{Y}$ is called the **error**. And the difference between $f, \hat{f}$ is called **reducible error**, because it can be potentially improved. The other part of the error which is due to the presence of the random noise ϵ, and cannot be reduced by our choice of the learning method, is called the **irreducible error**. We have

$$\begin{aligned}
\mathbb{E}[(Y - \hat{Y})^2] &= \mathbb{E}[(f(X) + \epsilon - \hat{f}(X))^2] \\
&= \mathbb{E}[(f(X) - \hat{f}(X))^2 + 2(f(X) - \hat{f}(X))\epsilon + \epsilon^2] \\
&= \mathbb{E}[(f(X) - \hat{f}(X))^2] + 2\mathbb{E}[(f(X) - \hat{f}(X))\epsilon] + \mathbb{E}[\epsilon^2] \\
&= (f(X) - \hat{f}(X))^2 + 2(f(X) - \hat{f}(X))\mathbb{E}[\epsilon] + \mathbb{E}[\epsilon^2] \\
&= (f(X) - \hat{f}(X))^2 + 0 + \mathbb{E}[(\epsilon - 0)^2] \\
&= \underbrace{(f(X) - \hat{f}(X))^2}_{\text{Reducible Error}} + \underbrace{\text{Var}[\epsilon]}_{\text{Irreducible Error}}.
\end{aligned}$$

1.2.2 Bias-Variance Decomposition

Suppose we have a data set consisting of n observations

$$(x_1, y_1), \ldots, (x_n, y_n)$$

for X, Y, where x_is can be vector too. Suppose we use this (training) data set in a learning method to find the estimate $\hat{f}$ for f. Then the **(training) Mean Squared Error (MSE)** is

$$\text{MSE} := \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

However, we are interested in the **test MSE**, i.e. the MSE of the model when it is applied to new test data.

Let us further assume that $f(\cdot) = F(\cdot, \theta_0)$ belongs to the class of functions

$$\mathcal{F} = \{F(\cdot, \theta) : \theta \in I\},$$

which is parametrized by θ in the parameter set I. Then we use a training data set to predict $\hat{f}(\cdot) = F(\cdot, \hat{\theta})$. Here $\hat{\theta}$ is the random variable which gives the parameter to obtain the prediction $\hat{f}$ when we use a particular training data set as input in our learning method. We assume that $\hat{\theta}, \epsilon$ are independent, since they are functions of two sources of randomness; ϵ is random due to unmeasured or unmeasurable variables, but $\hat{\theta}$ is random due to random selection of a training data set.

To simplify the notation we set $\mu = \mathbb{E}[\hat{Y}] = \mathbb{E}[\hat{f}(X)]$, i.e. μ is the average value of $\hat{Y}$ at X for different training data sets used as input in our learning method. Then the **expected test MSE** (i.e. the average of MSE w.r.t. all possible training data sets used as input in our learning method evaluated (tested) at X) can be computed as

$$\begin{aligned}
\mathbb{E}[(Y - \hat{Y})^2] &= \mathbb{E}\big[\big(f(X) + \epsilon - \hat{f}(X)\big)^2\big] \\
&= \mathbb{E}\big[(f(X) - \hat{f}(X))^2 + 2(f(X) - \hat{f}(X))\epsilon + \epsilon^2\big] \\
&= \mathbb{E}[(f(X) - \hat{f}(X))^2] + 2\mathbb{E}[(f(X) - \hat{f}(X))\epsilon] + \mathbb{E}[\epsilon^2] \\
&= \mathbb{E}[(f(X) - \hat{f}(X))^2] + 2\mathbb{E}[f(X) - \hat{f}(X)]\mathbb{E}[\epsilon] + \mathbb{E}[\epsilon^2] \\
&= \mathbb{E}[(f(X) - \mu + \mu - \hat{f}(X))^2] + 0 + \mathbb{E}[(\epsilon - 0)^2] \\
&= \mathbb{E}[(f(X) - \mu)^2] + \mathbb{E}[(\mu - \hat{f}(X))^2] \\
&\qquad + 2\mathbb{E}[(f(X) - \mu)(\mu - \hat{f}(X))] + \text{Var}[\epsilon] \\
&= (f(X) - \mu)^2 + \mathbb{E}[(\mu - \hat{f}(X))^2] \\
&\qquad + 2(f(X) - \mu)\mathbb{E}[\mu - \hat{f}(X)] + \text{Var}[\epsilon] \\
&= \underbrace{(f(X) - \mu)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}[(\mu - \hat{f}(X))^2]}_{\text{Variance}} + \underbrace{\text{Var}[\epsilon]}_{\text{Irreducible Error}}.
\end{aligned}$$

Here **variance** is the variance of the value of $\hat{Y}$ at X, and **bias** is the difference between the true value $f(X)$ with the mean value of $\hat{Y}$ at X. Hence we have shown that the reducible error is the sum of the variance and the square of bias.

In other words, suppose we have a large collection of training data sets, and we use each one of them to calculate an estimate $\hat{Y}$ for Y, using our learning method. Then this collection of estimates has a mean and a variance. The estimates are distributed around their mean, and their variability is quantified by their variance. The distance of the mean of all the estimates with the true value of $f(X)$ (which is the expected value of Y given X, since $\mathbb{E}[\epsilon] = 0$) is called the bias, and the variance of the estimates is still called the variance. Therefore bias measures the average distance that estimates have with $f(X)$. And variance measures how much the estimates are dispersed.

1.2.3 Error rates in Classification

Suppose Y is a qualitative variable, and its classes are represented by $\{1, 2, \ldots, K\}$. Let y_i be the ith observed value, and $\hat{y}_i$ be the predicted value for it. Then the **(training) error rate** is:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i),$$

where the indicator variable $I = 0$ if $y_i = \hat{y}_i$, and $I = 1$ if $y_i \neq \hat{y}_i$. Let y_0 be a new observed value, and $\hat{y}_0$ be its predicted value. Then the test error is

$$\mathbb{E}[I(y_0 \neq \hat{y}_0)].$$

Suppose we have predicted $\hat{y}_0 = j$. Then the **expected test error (ETE)** is

$$\begin{aligned}
\text{ETE} = \mathbb{E}[I(y_0 \neq \hat{y}_0)] &= \sum_{k=1}^{K} I(k \neq \hat{y}_0)\mathbb{P}(y_0 = k) \\
&= \sum_{k=1}^{K} I(k \neq j)\mathbb{P}(Y = k|X = x_0) \\
&= \sum_{k\neq j} \mathbb{P}(Y = k|X = x_0) = 1 - \mathbb{P}(Y = j|X = x_0).
\end{aligned}$$

Thus to minimize ETE we have to set $\hat{y}_0 = j$ such that $\mathbb{P}(Y = j|X = x_0)$ has the highest value among $\mathbb{P}(Y = k|X = x_0)$ for $k = 1, \ldots, K$. This is called the **Bayes classifier**. The **Bayes decision boundary** is the boundary between the regions $\{x_0 : \hat{y}_0 = j\}$ in the X-space, for different j. This boundary also determines the Bayes classifier.

Chapter 2

Linear Regression

Linear regression is a parametric learning method in which we assume that f is a linear function, i.e. we assume that

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

Then we need to estimate the coefficients $\beta_0, \ldots, \beta_p$. In **simple linear regression** we assume that $p = 1$, i.e. we have

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

The case of more than one predictor will be dealt with in **multiple linear regression**.

2.1 Simple Linear Regression

2.1.1 Least Squares Method

Suppose our data set consists of n observations

$$(x_1, y_1), \ldots, (x_n, y_n)$$

for X, Y. Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the predicted value for Y when $X = x_i$. Then $e_i = y_i - \hat{y}_i$ is called the ith **residual**. The **residual sum of squares (RSS)** is

$$\text{RSS} = e_1^2 + \cdots + e_n^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

In the **least squares** method we choose $\hat{\beta}_0, \hat{\beta}_1$ to minimize RSS.

To find the minimum of RSS we consider it as a function of $\hat{\beta}_0, \hat{\beta}_1$, and differentiate with respect to them. We get

$$0 = \frac{-1}{2}\frac{\partial}{\partial \hat{\beta}_0}\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = n(\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}),$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the **sample means** of x_i, y_i, respectively. Hence

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}.$$

Therefore

$$\begin{aligned}
0 = \frac{-1}{2}\frac{\partial}{\partial \hat{\beta}_1}\mathrm{RSS} &= \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
&= \sum_{i=1}^{n} x_i y_i - n\hat{\beta}_0 \bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} x_i y_i - n(\bar{y} - \hat{\beta}_1 \bar{x})\bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n}(x_i^2 - \bar{x}^2) \\
&= \sum_{i=1}^{n}\left[(x_i - \bar{x})(y_i - \bar{y}) + \bar{x} y_i + x_i \bar{y} - 2\bar{x}\bar{y}\right] \\
&\qquad - \hat{\beta}_1 \sum_{i=1}^{n}\left[(x_i - \bar{x})^2 + 2\bar{x}x_i - 2\bar{x}^2\right] \\
&= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2.
\end{aligned}$$

Thus

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

where $q_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ is the **sample covariance** of x_i, y_i, and $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the **sample variance** of x_i. We assume that s_x^2 is nonzero; otherwise x_is are constant.

Note that the second derivative of RSS is

$$D^2\mathrm{RSS} = \begin{bmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum x_i^2 \end{bmatrix}.$$

So $D^2\text{RSS}$ is positive definite, since its trace and determinant are positive:

$$\det(D^2\text{RSS}) = 4n\big(\sum x_i^2 - n\bar{x}^2\big) = 4n\sum(x_i - \bar{x})^2 = 4n(n-1)s_x^2.$$

Also note that RSS goes to infinity as $\hat{\beta}_0, \hat{\beta}_1$ go to infinity. Thus the above values for $\hat{\beta}_0, \hat{\beta}_1$ give the unique global minimum of RSS.

2.1.2 Unbiased Estimators

The reason that we use $n-1$ in the definition of s^2 instead of n, is that we want s^2 to be an **unbiased estimator** of (population) variance σ^2, i.e. the expected value of s^2 be equal to σ^2. To see this suppose $x_1, \ldots, x_n$ are independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2. First note that

$$\mathbb{E}[\bar{x}] = \mathbb{E}\Big[\frac{1}{n}\sum_{i\le n} x_i\Big] = \frac{1}{n}\sum_{i\le n}\mathbb{E}[x_i] = \frac{1}{n}n\mu = \mu.$$

Therefore the sample mean $\bar{x}$ is an unbiased estimator of the (population) mean μ. We also have

$$\begin{aligned}
\mathbb{E}[(\mu-\bar{x})^2] &= \frac{1}{n^2}\mathbb{E}\Big[\Big(n\mu - \sum_{i\le n} x_i\Big)^2\Big] = \frac{1}{n^2}\mathbb{E}\Big[\Big(\sum_{i\le n}(\mu - x_i)\Big)^2\Big] \\
&= \frac{1}{n^2}\mathbb{E}\Big[\sum_{i\le n}(\mu-x_i)^2 + \sum_{i\ne j}(\mu-x_i)(\mu-x_j)\Big] \\
&= \frac{1}{n^2}\sum_{i\le n}\mathbb{E}[(\mu-x_i)^2] + \frac{1}{n^2}\sum_{i\ne j}\mathbb{E}[(\mu-x_i)(\mu-x_j)] \\
&= \frac{n\sigma^2}{n^2} + \frac{1}{n^2}\sum_{i\ne j}\Big(\mathbb{E}[\mu-x_i]\mathbb{E}[\mu-x_j]\Big) \qquad (x_i, x_j \text{ are independent}) \\
&= \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n}.
\end{aligned}$$

Note that here we have shown that

$$\text{Var}(\bar{x}) = \mathbb{E}[(\bar{x}-\mu)^2] = \frac{\sigma^2}{n}.$$

Now we get

$$\begin{aligned}
\mathbb{E}\Big[\sum_{i\le n}(x_i-\bar{x})^2\Big] &= \mathbb{E}\Big[\sum_{i\le n}(x_i-\mu+\mu-\bar{x})^2\Big] \\
&= \mathbb{E}\Big[\sum_{i\le n}\big((x_i-\mu)^2+2(x_i-\mu)(\mu-\bar{x})+(\mu-\bar{x})^2\big)\Big] \\
&= \mathbb{E}\Big[\sum_{i\le n}(x_i-\mu)^2+2(\mu-\bar{x})\sum_{i\le n}(x_i-\mu)+n(\mu-\bar{x})^2\Big] \\
&= \mathbb{E}\Big[\sum_{i\le n}(x_i-\mu)^2+2(\mu-\bar{x})n(\bar{x}-\mu)+n(\mu-\bar{x})^2\Big] \\
&= \mathbb{E}\Big[\sum_{i\le n}(x_i-\mu)^2-n(\mu-\bar{x})^2\Big] \\
&= \sum_{i\le n}\mathbb{E}[(x_i-\mu)^2]-n\mathbb{E}[(\mu-\bar{x})^2] \\
&= n\sigma^2-n\frac{\sigma^2}{n}=(n-1)\sigma^2.
\end{aligned}$$

Therefore $\mathbb{E}[s^2]=\sigma^2$, as desired.

On the other hand, notice that in general s is not an unbiased estimator of the standard deviation σ, since in general $\mathbb{E}[s]=\mathbb{E}[\sqrt{s^2}]\neq\sqrt{\mathbb{E}[s^2]}=\sigma$.

Next let us show that $\hat{\beta}_0,\hat{\beta}_1$ are unbiased estimators of β_0,β_1, respectively. First note that we have $y_i=\beta_0+\beta_1x_i+\epsilon_i$, where ϵ_i,x_i are independent, and $\mathbb{E}[\epsilon_i]=0$ (we also assume that $x_i,x_j,\epsilon_i,\epsilon_j$ are all independent for $i\neq j$). Hence we have

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1] &= \mathbb{E}\Big[\frac{\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^n(x_i-\bar{x})^2}\Big] \\
&= \mathbb{E}\Big[\frac{\sum_{i=1}^n(x_i-\bar{x})(\beta_0+\beta_1x_i+\epsilon_i-\beta_0-\beta_1\bar{x}-\bar{\epsilon})}{\sum_{i=1}^n(x_i-\bar{x})^2}\Big] \\
&= \mathbb{E}\Big[\frac{\beta_1\sum_{i=1}^n(x_i-\bar{x})^2+\sum_{i=1}^n(x_i-\bar{x})(\epsilon_i-\bar{\epsilon})}{\sum_{i=1}^n(x_i-\bar{x})^2}\Big] \\
&= \beta_1+\sum_{i=1}^n\mathbb{E}\Big[\frac{(x_i-\bar{x})}{\sum_{i=1}^n(x_i-\bar{x})^2}(\epsilon_i-\bar{\epsilon})\Big] \\
&= \beta_1+\sum_{i=1}^n\mathbb{E}\Big[\frac{(x_i-\bar{x})}{\sum_{i=1}^n(x_i-\bar{x})^2}\Big]\mathbb{E}[\epsilon_i-\bar{\epsilon}]=\beta_1+0=\beta_1.
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{y} - \hat{\beta}_1\bar{x}] &= \mathbb{E}\Big[\bar{y} - \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\bar{x}\Big] \\
&= \mathbb{E}\Big[\bar{y} - \frac{\beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (x_i - \bar{x})(\epsilon_i - \bar{\epsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2}\bar{x}\Big] \\
&= \mathbb{E}[\bar{y} - \beta_1\bar{x}] - \sum_{i=1}^n \mathbb{E}\Big[\frac{(x_i - \bar{x})\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}(\epsilon_i - \bar{\epsilon})\Big] \\
&= \mu_y - \beta_1\mu_x - \sum_{i=1}^n \mathbb{E}\Big[\frac{(x_i - \bar{x})\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\Big]\mathbb{E}[\epsilon_i - \bar{\epsilon}] \\
&= \beta_0 - 0 = \beta_0.
\end{aligned}$$

2.1.3 t-statistic and Hypothesis tests

In the last section we computed the expected values of $\hat{\beta}_0, \hat{\beta}_1$. We can also compute the variances of $\hat{\beta}_0, \hat{\beta}_1$ similarly, but it is easier to compute them using matrix algebra, and we will do so in Section 2.2.2. The square root of the variance of an estimate is called its **standard error**. Let us denote the standard errors of $\hat{\beta}_0, \hat{\beta}_1$ by $\mathrm{SE}(\hat{\beta}_0), \mathrm{SE}(\hat{\beta}_1)$ respectively.

Now suppose we want to see whether there really is a relationship between X, Y. In other words, we want to see whether the observed data provides statistically significant evidence to support that X, Y are related. To do this we can perform a **hypothesis test**. We start with a **null hypothesis**. Here our null hypothesis is that there is no relation between X, Y, which mathematically can be expressed as

$$H_0 : \beta_1 = 0.$$

The idea of the hypothesis test is that if the data shows that the null hypothesis results in a statistically improbable conclusion, then we can reject the null hypothesis, and accept its negation, which is called the **alternative hypothesis**. Thus hypothesis test can be considered a statistical form of reductio ad absurdum. In our case, the alternative hypothesis is that there is some relation between X, Y, or more formally

$$H_a : \beta_1 \neq 0.$$

To perform hypothesis test for the null hypothesis $H_0 : \beta_1 = 0$ we can use the **t-statistic**

$$t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

which measures the distance of $\hat{\beta}_1$ to 0 in the units of the standard error of $\hat{\beta}_1$. The distribution of the t-statistic can be computed, and is called **t-distribution**.

The t-distribution is similar to the normal distribution, and is concentrated around zero. It can be shown that if H_0 is true then it is unlikely that the t-statistic has a large value. So if we observe a large t-statistic, then we can reject H_0. To make this reasoning quantitative, let T be a random variable with t-distribution. Then the **p-value** of the observed t-statistic is

$$p = \mathbb{P}(|T| \geq |t| \,|\, H_0),$$

i.e. the p-value is the probability of observing t or less probable values than t, provided that H_0 is true. Therefore if the p-value of t is small enough, we can conclude that the null hypothesis H_0 implies an improbable result, so it can be rejected.

2.1.4 R^2 Statistic

The R^2 statistic, also called the **coefficient of determination**, is

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

where $\text{TSS} = \sum(y_i - \bar{y})^2$ is the **total sum of squares**. Intuitively, TSS is the total variation in the observed values of Y, and RSS is the variation in Y after we perform the regression. Thus we can regard RSS as the variation in Y which is unexplained by the linear regression. Hence $\text{TSS} - \text{RSS}$ is the amount of variation in Y which is explained by the linear regression, and therefore R^2 is the proportion of the variance in the response which is explained by the linear regression. So if R^2 is close to 1, the linear model is a good approximation of the true relationship between X, Y. There is an alternative way to see this fact, which we present in the next paragraph.

Recall that the **sample correlation** is

$$\text{Cor}(X, Y) = r_{xy} = \frac{q_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

We have

$$\begin{aligned} R^2 &= \frac{\text{TSS} - \text{RSS}}{\text{TSS}} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \bar{y} + \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{i=1}^n \left[(y_i-\bar{y})^2 - (y_i - \bar{y} + \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i)^2\right]}{\sum_{i=1}^n (y_i-\bar{y})^2} && (\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}) \\
&= \frac{\sum_{i=1}^n \left[2(y_i-\bar{y})\hat{\beta}_1(x_i-\bar{x}) - \hat{\beta}_1^2(x_i-\bar{x})^2\right]}{\sum_{i=1}^n (y_i-\bar{y})^2} \\
&= \frac{2\hat{\beta}_1 \sum_{i=1}^n (y_i-\bar{y})(x_i-\bar{x}) - \hat{\beta}_1^2 \sum_{i=1}^n (x_i-\bar{x})^2}{\sum_{i=1}^n (y_i-\bar{y})^2} \\
&= \frac{2\hat{\beta}_1 q_{xy} - \hat{\beta}_1^2 s_x^2}{s_y^2} = \frac{2\frac{q_{xy}}{s_x^2} q_{xy} - \frac{q_{xy}^2}{s_x^4} s_x^2}{s_y^2} = \frac{q_{xy}^2}{s_x^2 s_y^2} = \mathrm{Cor}^2(X,Y).
\end{aligned}
$$

It is easy to see that $\mathrm{Cor}(\hat{Y}, Y) = \mathrm{Cor}(\hat{\beta}_0 + \hat{\beta}_1 X, Y) = \mathrm{Cor}(X, Y)$. Hence we also have $R^2 = \mathrm{Cor}^2(\hat{Y}, Y)$.

2.2 Multiple Linear Regression

2.2.1 Estimating the Coefficients

Suppose that we have

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

Let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ be estimates for the above coefficients. Then our prediction for Y is given by the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p.$$

We want to find $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize the sum of squared residuals

$$\mathrm{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2,$$

i.e. we want to use the method of least squares. Set

$$
\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}, \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}.
$$

Then we have $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, and

$$\mathrm{RSS} = \mathrm{RSS}(\hat{\boldsymbol{\beta}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\intercal (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

We want RSS to have its minimum at $\hat{\boldsymbol{\beta}}$. Let $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$ be an arbitrary vector. Then the directional derivative of RSS at $\hat{\boldsymbol{\beta}}$ in the direction of $\boldsymbol{\gamma}$ must be zero. Hence we must have

$$\begin{aligned}
0 &= \frac{d}{dt}\Big|_{t=0} \mathrm{RSS}(\hat{\boldsymbol{\beta}} + t\boldsymbol{\gamma}) \\
&= \frac{d}{dt}\Big|_{t=0} \big((\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - t\mathbf{X}\boldsymbol{\gamma})^{\intercal}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - t\mathbf{X}\boldsymbol{\gamma})\big) \\
&= \big(\frac{d}{dt}\Big|_{t=0}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - t\mathbf{X}\boldsymbol{\gamma})^{\intercal}\big)(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\intercal}\big(\frac{d}{dt}\Big|_{t=0}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - t\mathbf{X}\boldsymbol{\gamma})\big) \\
&= (-\mathbf{X}\boldsymbol{\gamma})^{\intercal}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\intercal}(-\mathbf{X}\boldsymbol{\gamma}) \\
&= -\boldsymbol{\gamma}^{\intercal}\mathbf{X}^{\intercal}\mathbf{y} + \boldsymbol{\gamma}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y}^{\intercal}\mathbf{X}\boldsymbol{\gamma} + \hat{\boldsymbol{\beta}}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\boldsymbol{\gamma} \\
&= -2\boldsymbol{\gamma}^{\intercal}\mathbf{X}^{\intercal}\mathbf{y} + 2\boldsymbol{\gamma}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\hat{\boldsymbol{\beta}} \qquad \text{(since } a^{\intercal} = a \text{ for a } 1 \times 1 \text{ matrix)} \\
&= 2\boldsymbol{\gamma}^{\intercal}\big(-\mathbf{X}^{\intercal}\mathbf{y} + \mathbf{X}^{\intercal}\mathbf{X}\hat{\boldsymbol{\beta}}\big).
\end{aligned}$$

Therefore

$$\mathbf{X}^{\intercal}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\intercal}\mathbf{y},$$

since $\boldsymbol{\gamma}$ is arbitrary, and $\mathbf{X}^{\intercal}\mathbf{y} - \mathbf{X}^{\intercal}\mathbf{X}\hat{\boldsymbol{\beta}}$ is also an element of $\mathbb{R}^{p+1}$. If $\mathbf{X}^{\intercal}\mathbf{X}$ is invertible, which is the case if the columns of $\mathbf{X}$ are linearly independent, then we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{y}.$$

Let us present another way for finding $\hat{\boldsymbol{\beta}}$. Note that in general for every $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$ we have

$$0 = \langle 0, \boldsymbol{\gamma}\rangle = \langle \mathbf{X}^{\intercal}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \boldsymbol{\gamma}\rangle = \langle \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}, \mathbf{X}\boldsymbol{\gamma}\rangle.$$

In other words, $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ must be orthogonal to the image of the linear map $\cdot \mapsto \mathbf{X}\cdot$, which is a linear subspace of $\mathbb{R}^n$. For this to happen, $\mathbf{X}\hat{\boldsymbol{\beta}}$ must be the **orthogonal projection** of $\mathbf{y}$ onto this subspace; and in this case we also have

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\| \leq \|\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}\|,$$

for every $\boldsymbol{\gamma}$, which is the desired property of $\hat{\boldsymbol{\beta}}$. We should also note that there is always a $\hat{\boldsymbol{\beta}}$ that satisfies the above inequality, and $\mathbf{X}\hat{\boldsymbol{\beta}}$ is uniquely determined as the orthogonal projection of $\mathbf{y}$; but when $\mathbf{X}^{\intercal}\mathbf{X}$ is non-invertible, $\hat{\boldsymbol{\beta}}$ is not unique.

In addition, note that we actually have

$$\begin{aligned}
D_{\boldsymbol{\gamma}}\mathrm{RSS}(\hat{\boldsymbol{\beta}}) = \frac{d}{dt}\Big|_{t=0} \mathrm{RSS}(\hat{\boldsymbol{\beta}} + t\boldsymbol{\gamma}) &= 2\boldsymbol{\gamma}^{\intercal}\big(\mathbf{X}^{\intercal}\mathbf{y} - \mathbf{X}^{\intercal}\mathbf{X}\hat{\boldsymbol{\beta}}\big) \\
&= 2\big(\mathbf{y}^{\intercal}\mathbf{X} - \hat{\boldsymbol{\beta}}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\big)\boldsymbol{\gamma}.
\end{aligned}$$

Hence $D\mathrm{RSS}(\hat{\boldsymbol{\beta}}) = 2\big(\mathbf{y}^\intercal\mathbf{X} - \hat{\boldsymbol{\beta}}^\intercal\mathbf{X}^\intercal\mathbf{X}\big)$. Therefore

$$D^2\mathrm{RSS}(\hat{\boldsymbol{\beta}}) = 2\mathbf{X}^\intercal\mathbf{X}$$

is positive definite. Also note that RSS goes to infinity as $\hat{\boldsymbol{\beta}}$ goes to infinity. Thus the above value for $\hat{\boldsymbol{\beta}}$ gives the unique global minimum of RSS.

It is also easy to see that

$$\begin{aligned} 0 = \frac{\partial}{\partial\hat{\beta}_0}\mathrm{RSS} &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) \\ &= -2n(\bar{y} - \hat{\beta}_0 - \hat{\beta}_1\bar{x}_1 - \cdots - \hat{\beta}_p\bar{x}_p). \end{aligned}$$

Hence $\bar{\hat{y}} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \cdots + \hat{\beta}_p\bar{x}_p = \bar{y}$. In other words, the linear function given by the least squares method passes through the sample means of the data.

2.2.2 Standard Errors of the Coefficient Estimates

Suppose the true relationship between $\mathbf{X}, \mathbf{y}$ is

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

where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\intercal$ is a vector of i.i.d random variables with 0 mean and variance σ_ϵ^2. Thus its covariance matrix is

$$\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal] = \begin{bmatrix} \mathbb{E}[\epsilon_1^2] & \mathbb{E}[\epsilon_1\epsilon_2] & \cdots & \mathbb{E}[\epsilon_1\epsilon_n] \\ \mathbb{E}[\epsilon_1\epsilon_2] & \mathbb{E}[\epsilon_2^2] & \cdots & \mathbb{E}[\epsilon_2\epsilon_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[\epsilon_1\epsilon_n] & \mathbb{E}[\epsilon_2\epsilon_n] & \cdots & \mathbb{E}[\epsilon_n^2] \end{bmatrix} = \sigma_\epsilon^2 I,$$

where I is the identity matrix (note that all covariances are 0 due to independence). We also assume that $\mathbf{X}, \boldsymbol{\epsilon}$ are independent. Now we have

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon} = \boldsymbol{\beta} + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon}. \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}\big[\boldsymbol{\beta} + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon}\big] \\ &= \boldsymbol{\beta} + \mathbb{E}\big[(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\big]\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{\beta} + 0 = \boldsymbol{\beta}. \end{aligned}$$

Therefore $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. On the other hand, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}
\mathbb{E}\big[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\intercal}\big] &= \mathbb{E}\big[\big((\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{y}-\boldsymbol{\beta}\big)\big((\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{y}-\boldsymbol{\beta}\big)^{\intercal}\big] \\
&= \mathbb{E}\big[\big((\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\boldsymbol{\epsilon}\big)\big((\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\boldsymbol{\epsilon}\big)^{\intercal}\big] \\
&= (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\intercal}]\mathbf{X}(\mathbf{X}^{\intercal}\mathbf{X})^{-1} \\
&= (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\sigma_{\epsilon}^{2}I\mathbf{X}(\mathbf{X}^{\intercal}\mathbf{X})^{-1} \\
&= \sigma_{\epsilon}^{2}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{X}(\mathbf{X}^{\intercal}\mathbf{X})^{-1} = \sigma_{\epsilon}^{2}(\mathbf{X}^{\intercal}\mathbf{X})^{-1},
\end{aligned}$$

provided that $\mathbf{X}$ is not random (or we compute the conditional expectation under the condition that $\mathbf{X}$ is constant).

Suppose $p=1$. Then we have (for simplicity, we write x_i instead of x_{i1})

$$\mathbf{X}^{\intercal}\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}.$$

Hence

$$\begin{aligned}
(\mathbf{X}^{\intercal}\mathbf{X})^{-1} &= \frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \\
&= \frac{1}{n\sum (x_i-\bar{x})^2}\begin{bmatrix} \sum (x_i-\bar{x})^2 + n\bar{x}^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}.
\end{aligned}$$

Note that here we have used the formula

$$\sum (x_i-\bar{x})^2 = \sum x_i^2 - \frac{1}{n}\big(\sum x_i\big)^2,$$

which is a manifestation of the formula $\mathbb{E}[(Z-\mathbb{E}[Z])^2] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2$, for a random variable Z. Hence we obtain the following formulas for the **standard errors** of $\hat{\beta}_0, \hat{\beta}_1$

$$\begin{aligned}
\mathrm{SE}(\hat{\beta}_0)^2 &= \mathrm{Var}(\hat{\beta}_0) = \sigma_{\epsilon}^2\Big[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\Big], \\
\mathrm{SE}(\hat{\beta}_1)^2 &= \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma_{\epsilon}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}.
\end{aligned}$$

To compute the standard errors of the coefficients when p is arbitrary we need some tools from linear algebra. It is easy to see that the inverse of a block matrix

is given by any of the following two formulas

$$\begin{aligned}\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} &= \begin{bmatrix} (A-BD^{-1}C)^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1} \\ -D^{-1}C(A-BD^{-1}C)^{-1} & (D-CA^{-1}B)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} A^{-1}+A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1} \\ -(D-CA^{-1}B)^{-1}CA^{-1} & (D-CA^{-1}B)^{-1} \end{bmatrix},\end{aligned}$$

provided that all the inverses exist and all the products are defined. (For the proof we can just multiply the proposed inverses and the original block matrix.) The matrices $(A-BD^{-1}C), (D-CA^{-1}B)$ are called the **Schur complements** of D, A, respectively. Note that we also have

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & 0 \\ C & I \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & D-CA^{-1}B \end{bmatrix}.$$

Therefore the determinant of the block matrix equals $\det(A)\det(D-CA^{-1}B)$. So if A and the block matrix are invertible then the Schur complement of A is invertible too.

Now in general we can write the matrix $\mathbf{X}$ as $\mathbf{X} = \begin{bmatrix} \mathbf{1} & \tilde{\mathbf{X}} \end{bmatrix}$, where $\mathbf{1} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}^\intercal \in \mathbb{R}^n$. Hence

$$\mathbf{X}^\intercal\mathbf{X} = \begin{bmatrix} \mathbf{1}^\intercal \\ \tilde{\mathbf{X}}^\intercal \end{bmatrix} \begin{bmatrix} \mathbf{1} & \tilde{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} n & \mathbf{1}^\intercal\tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^\intercal\mathbf{1} & \tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}} \end{bmatrix} = \begin{bmatrix} n & n\bar{\boldsymbol{x}}^\intercal \\ n\bar{\boldsymbol{x}} & \tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}} \end{bmatrix},$$

where $\bar{\boldsymbol{x}} = \begin{bmatrix} \bar{x}_1 & \dots & \bar{x}_p \end{bmatrix}^\intercal$. Thus

$$\begin{aligned}(\mathbf{X}^\intercal\mathbf{X})^{-1} &= \begin{bmatrix} n^{-1}+n^{-1}n\bar{\boldsymbol{x}}^\intercal(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}-n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal)^{-1}n\bar{\boldsymbol{x}}n^{-1} & -n^{-1}n\bar{\boldsymbol{x}}^\intercal(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}-n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal)^{-1} \\ -(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}-n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal)^{-1}n\bar{\boldsymbol{x}}n^{-1} & (\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}-n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal)^{-1} \end{bmatrix} \\ &= \frac{1}{n-1}\begin{bmatrix} n^{-1}(n-1)+\bar{\boldsymbol{x}}^\intercal Q^{-1}\bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}^\intercal Q^{-1} \\ -Q^{-1}\bar{\boldsymbol{x}} & Q^{-1} \end{bmatrix},\end{aligned}$$

where $Q = (n-1)^{-1}(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal)$ is the matrix of sample covariances, since

$$(\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^\intercal)_{ij} = \sum_{k=1}^{n} x_{ki}x_{kj} - n\bar{x}_i\bar{x}_j = \sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j).$$

Therefore

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma_\epsilon^2\Big[\frac{1}{n} + \frac{1}{n-1}\bar{\boldsymbol{x}}^\intercal Q^{-1}\bar{\boldsymbol{x}}\Big], \qquad \mathrm{SE}(\hat{\beta}_j)^2 = \frac{1}{n-1}\sigma_\epsilon^2(Q^{-1})_{jj}.$$

2.2.3 Estimating the Irreducible Error

Let us first compute $\mathbb{E}[\text{RSS}]$. We have

$$\begin{aligned}
\mathbb{E}[\text{RSS}] &= \mathbb{E}[(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y})^\intercal(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y})] \\
&= \mathbb{E}[(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\epsilon})^\intercal(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\epsilon})] \\
&= \mathbb{E}\big[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\intercal\mathbf{X}^\intercal\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\epsilon}^\intercal\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\intercal\mathbf{X}^\intercal\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\intercal\boldsymbol{\epsilon}\big] \\
&= \mathbb{E}\big[\big((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon}\big)^\intercal\mathbf{X}^\intercal\mathbf{X}\big((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon}\big) \\
&\qquad\qquad - \boldsymbol{\epsilon}^\intercal\mathbf{X}\big((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon}\big) - \big((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon}\big)^\intercal\mathbf{X}^\intercal\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\intercal\boldsymbol{\epsilon}\big] \\
&= \mathbb{E}[-\boldsymbol{\epsilon}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\intercal\boldsymbol{\epsilon}].
\end{aligned}$$

Now note that for every vector $\boldsymbol{v}$ the scalar $\boldsymbol{v}^\intercal\boldsymbol{v}$ equals the trace of the matrix $\boldsymbol{v}\boldsymbol{v}^\intercal$. Hence in particular we have

$$\mathbb{E}[\boldsymbol{\epsilon}^\intercal\boldsymbol{\epsilon}] = \mathbb{E}[\text{tr}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal)] = \text{tr}(\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal]) = \text{tr}(\sigma_\epsilon^2 I) = n\sigma_\epsilon^2.$$

Note that trace and $\mathbb{E}$ commute, since trace is linear.

On the other hand, we know that $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ is a symmetric positive definite matrix. Thus we have $(\mathbf{X}^\intercal\mathbf{X})^{-1} = \mathbf{S}^2$ for some symmetric positive definite matrix $\mathbf{S}$. Now note that $\mathbf{S}\mathbf{X}^\intercal\boldsymbol{\epsilon}$ is a $(p+1)$-dimensional random vector. The covariance matrix of this vector is equal to

$$\begin{aligned}
\mathbb{E}[\mathbf{S}\mathbf{X}^\intercal\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal\mathbf{X}\mathbf{S}] &= \mathbf{S}\mathbf{X}^\intercal\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal]\mathbf{X}\mathbf{S} \qquad\qquad \text{(since } \mathbf{X}\text{,}\mathbf{S} \text{ are constant)} \\
&= \mathbf{S}\mathbf{X}^\intercal\sigma_\epsilon^2 I\mathbf{X}\mathbf{S} = \sigma_\epsilon^2\mathbf{S}\mathbf{X}^\intercal\mathbf{X}\mathbf{S} = \sigma_\epsilon^2\mathbf{S}(\mathbf{S}^2)^{-1}\mathbf{S} = \sigma_\epsilon^2 I.
\end{aligned}$$

Note that here I is the $p+1$ by $p+1$ identity matrix. Now we have

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\epsilon}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon}] &= \mathbb{E}[\boldsymbol{\epsilon}^\intercal\mathbf{X}\mathbf{S}\mathbf{S}\mathbf{X}^\intercal\boldsymbol{\epsilon}] \\
&= \mathbb{E}[(\boldsymbol{\epsilon}^\intercal\mathbf{X}\mathbf{S})(\mathbf{S}\mathbf{X}^\intercal\boldsymbol{\epsilon})] \\
&= \mathbb{E}[(\mathbf{S}\mathbf{X}^\intercal\boldsymbol{\epsilon})^\intercal(\mathbf{S}\mathbf{X}^\intercal\boldsymbol{\epsilon})] \\
&= \mathbb{E}[\text{tr}((\mathbf{S}\mathbf{X}^\intercal\boldsymbol{\epsilon})(\mathbf{S}\mathbf{X}^\intercal\boldsymbol{\epsilon})^\intercal)] \\
&= \mathbb{E}[\text{tr}(\mathbf{S}\mathbf{X}^\intercal\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal\mathbf{X}\mathbf{S})] \\
&= \text{tr}(\mathbb{E}[\mathbf{S}\mathbf{X}^\intercal\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal\mathbf{X}\mathbf{S}]) = \text{tr}(\sigma_\epsilon^2 I) = (p+1)\sigma_\epsilon^2.
\end{aligned}$$

Hence we get

$$\mathbb{E}[\text{RSS}] = \mathbb{E}[-\boldsymbol{\epsilon}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^\intercal\boldsymbol{\epsilon}] = (n-p-1)\sigma_\epsilon^2.$$

Thus the square of **residual standard error**

$$\text{RSE}^2 = \frac{\text{RSS}}{n-p-1}$$

is an unbiased estimator of σ_ϵ^2.

2.2.4 F-Statistic

Suppose $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$. To test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0,$$

we use the **F-statistic**

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

where $\text{TSS} = \sum(y_i - \bar{y})^2$ and $\text{RSS} = \sum(y_i - \hat{y}_i)^2$.

More generally, if we want to test the null hypothesis

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0,$$

we use the F-statistic

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)}.$$

Here RSS_0 is the residual sum of squares when we fit a second model using the first $p - q$ variables, i.e. RSS_0 corresponds to a smaller model with predictors $X_1, \ldots, X_{p-q}$; and RSS corresponds to the larger original model with all the predictors $X_1, \ldots, X_{p-q}, \ldots, X_p$. (Note that TSS is the RSS of a fitted model which has no predictors, and has only the intercept term. So the previous formula for F is a special case of the above formula.)

The F statistic is a normalized measure of change in residual sum of squares (which can be thought of as the total error of prediction in the linear model) per added predictor, when we extend the model with predictors $X_1, \ldots, X_{p-q}$ by adding the predictors $X_{p-q+1}, \ldots, X_p$; and we have computed this measure (the F statistic) in the units of our estimate of σ_ϵ^2, which is itself a measure of the noise in the response Y.

When $\text{RSS}_0 - \text{RSS}$ is small (which hints that H_0 might be true), the smaller model with $p - q$ features explains the response nearly as good as the larger model with p features. Therefore it might be the case that adding the last q variables to the model does not improve our estimation of Y significantly; so we cannot reject the null hypothesis H_0.

We have seen that $\mathbb{E}[\text{RSS}/(n-p-1)] = \sigma_\epsilon^2$. Let

$$Z = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Then Z, ϵ are uncorrelated, since

$$\text{Cov}(Z, \epsilon) = \text{Cov}(\beta_0, \epsilon) + \beta_1 \text{Cov}(X_1, \epsilon) + \cdots + \beta_p \text{Cov}(X_p, \epsilon) = 0,$$

because X_i, ϵ are independent for each i, and β_0 is a constant. Now we have

$$\sigma_Y^2 = \sigma_{Z+\epsilon}^2 = \sigma_Z^2 + 2\text{Cov}(Z, \epsilon) + \sigma_\epsilon^2 = \sigma_Z^2 + \sigma_\epsilon^2.$$

Hence we also have

$$\mathbb{E}[\mathrm{TSS}] = \mathbb{E}[(n-1)s_y^2] = (n-1)\sigma_Y^2 = (n-1)(\sigma_Z^2 + \sigma_\epsilon^2) \geq (n-1)\sigma_\epsilon^2.$$

In particular, when $q = p$ and H_0 is true, we have $\mathbb{E}[(\mathrm{TSS} - \mathrm{RSS})/p] = \sigma_\epsilon^2$, since Z is constant in this case; otherwise we have $\mathbb{E}[(\mathrm{TSS} - \mathrm{RSS})/p] \geq \sigma_\epsilon^2$. Therefore when $q = p$ and H_0 is true we expect the F-statistic be close to 1. Otherwise we expect F to be greater than 1. In general, under suitable assumptions, the distribution of F is known, and is called **F-distribution**. Using F-distribution we can assign p-value to F, and perform hypothesis testing.

Let us inspect the F-statistic more closely. Let $\hat{\mathbf{y}}$ and $\tilde{\mathbf{y}}$ be the estimates of $\mathbf{y}$ using the larger and the smaller models respectively. We know that $\tilde{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ on the subspace W which is spanned by the vectors $\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_{p-q}$, and $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ on the subspace V which is spanned by

$$\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_{p-q}, \ldots, \mathbf{x}_p.$$

Obviously we have $\hat{\mathbf{y}} \in V$ and $\tilde{\mathbf{y}} \in W \subset V$. Also note that due to the properties of orthogonal projections we have

$$\mathbf{y} - \tilde{\mathbf{y}} \in W^\perp, \qquad \mathbf{y} - \hat{\mathbf{y}} \in V^\perp.$$

Therefore we get

$$\begin{aligned} \mathrm{RSS}_0 = \|\mathbf{y} - \tilde{\mathbf{y}}\|^2 &= \|\mathbf{y} - \hat{\mathbf{y}} + \hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2 \\ &= \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2 \qquad (\text{since } \mathbf{y} - \hat{\mathbf{y}} \perp \hat{\mathbf{y}} - \tilde{\mathbf{y}}) \\ &= \mathrm{RSS} + \|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2. \end{aligned}$$

An immediate consequence of the above equation is that RSS decreases as we add more predictors to the model. Intuitively, this happens because RSS measures the prediction error using the training data, and with more predictors we have more degrees of freedom, so the training error becomes smaller.

Let us find a geometric description of $\hat{\mathbf{y}} - \tilde{\mathbf{y}}$. Let $U = V \cap W^\perp$ be the orthogonal complement of W in V. Then note that

$$\mathbf{y} - (\hat{\mathbf{y}} - \tilde{\mathbf{y}}) = (\mathbf{y} - \hat{\mathbf{y}}) + \tilde{\mathbf{y}} \in U^\perp,$$

since $\mathbf{y} - \hat{\mathbf{y}} \in V^\perp \subset U^\perp$, and $\tilde{\mathbf{y}} \in W = (W^\perp)^\perp \subset U^\perp$. Hence $\hat{\mathbf{y}} - \tilde{\mathbf{y}}$ must be the orthogonal projection of $\mathbf{y}$ onto the subspace U. Therefore we have the following formula for the F-statistic

$$\begin{aligned} F = \frac{\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2/q}{\|\mathbf{y} - \hat{\mathbf{y}}\|^2/(n-p-1)} &= \frac{n-p-1}{q} \frac{\|P_U \mathbf{y}\|^2}{\|\mathbf{y} - P_V \mathbf{y}\|^2} \\ &= \frac{n-p-1}{q} \frac{\|P_{V \cap W^\perp} \mathbf{y}\|^2}{\|P_{V^\perp} \mathbf{y}\|^2}, \end{aligned}$$

where P denotes the orthogonal projection operator on a subspace.

2.2.5 F-Statistic versus t-statistic

For each predictor we have a t-statistic. It can be shown that the square of this t-statistic is the F-statistic corresponding to omitting that variable from the model, i.e. when $q = 1$. To see this, note that we have

$$\mathbf{X} = \begin{bmatrix}\mathbf{X}_{\tilde{p}} & \mathbf{x}_p\end{bmatrix} = \begin{bmatrix}\mathbf{1} & \mathbf{x}_1 & \dots & \mathbf{x}_p\end{bmatrix},$$

where $\mathbf{x}_j = \begin{bmatrix}x_{1j} & \dots & x_{nj}\end{bmatrix}^\intercal$, and $\mathbf{1} = \begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix}^\intercal \in \mathbb{R}^n$. Let

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal, \qquad \mathbf{H}_{\tilde{p}} = \mathbf{X}_{\tilde{p}}(\mathbf{X}_{\tilde{p}}^\intercal\mathbf{X}_{\tilde{p}})^{-1}\mathbf{X}_{\tilde{p}}^\intercal.$$

Then $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ and $\tilde{\mathbf{y}} = \mathbf{H}_{\tilde{p}}\mathbf{y}$ are respectively the predictions of Y based on all the predictors and the first $p-1$ predictors. Let V_j be the linear subspace spanned by the first $j+1$ columns of $\mathbf{X}$. We have $V_j \subset V_{j+1}$, and since we assume that the columns of $\mathbf{X}$ are linearly independent, we have $\dim V_j = j+1$. We also know that the predictions based on least squares method actually give us the orthogonal projections on the subspace spanned by the columns. Hence $\mathbf{H}, \mathbf{H}_{\tilde{p}}$ are the **projection (hat) matrices** on the subspaces V_p, V_{p-1}, respectively.

Let $\tilde{\mathbf{x}}_p = \mathbf{H}_{\tilde{p}}\mathbf{x}_p$ be the projection of $\mathbf{x}_p$ on V_{p-1}, i.e. the prediction of X_p using the other predictors. Then we have

$$\begin{aligned}\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} &= \hat{\beta}_0\mathbf{1} + \hat{\beta}_1\mathbf{x}_1 + \cdots + \hat{\beta}_{p-1}\mathbf{x}_{p-1} + \hat{\beta}_p\mathbf{x}_p \\ &= \hat{\beta}_0\mathbf{1} + \hat{\beta}_1\mathbf{x}_1 + \cdots + \hat{\beta}_{p-1}\mathbf{x}_{p-1} + \hat{\beta}_p\tilde{\mathbf{x}}_p + \hat{\beta}_p(\mathbf{x}_p - \tilde{\mathbf{x}}_p).\end{aligned}$$

But note that $\mathbf{x}_p - \tilde{\mathbf{x}}_p$ is orthogonal to V_{p-1}, and

$$\mathbf{v} = \hat{\beta}_0\mathbf{1} + \hat{\beta}_1\mathbf{x}_1 + \cdots + \hat{\beta}_{p-1}\mathbf{x}_{p-1} + \hat{\beta}_p\tilde{\mathbf{x}}_p \in V_{p-1}.$$

Now we have $\mathbf{y} - \mathbf{v} = \mathbf{y} - \hat{\mathbf{y}} + \hat{\beta}_p(\mathbf{x}_p - \tilde{\mathbf{x}}_p)$. But $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to $V_p \supset V_{p-1}$, and $\mathbf{x}_p - \tilde{\mathbf{x}}_p$ is orthogonal to V_{p-1}; so $\mathbf{y} - \mathbf{v}$ is orthogonal to V_{p-1}. Hence $\mathbf{v}$ must be the orthogonal projection of $\mathbf{y}$ on V_{p-1}, i.e. $\mathbf{v} = \tilde{\mathbf{y}} = \mathbf{H}_{\tilde{p}}\mathbf{y}$. Thus we have

$$\mathrm{RSS}_0 = \|\mathbf{y} - \tilde{\mathbf{y}}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \hat{\beta}_p^2\|\mathbf{x}_p - \tilde{\mathbf{x}}_p\|^2 = \mathrm{RSS} + \hat{\beta}_p^2\|\mathbf{x}_p - \tilde{\mathbf{x}}_p\|^2,$$

since $\mathbf{x}_p - \tilde{\mathbf{x}}_p$ belongs to V_p, and $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to V_p. Again, we see that RSS must decrease (or remain constant) when we add new predictors.

On the other hand we have

$$\mathbf{X}^\intercal\mathbf{X} = \begin{bmatrix}\mathbf{X}_{\tilde{p}}^\intercal \\ \mathbf{x}_p^\intercal\end{bmatrix}\begin{bmatrix}\mathbf{X}_{\tilde{p}} & \mathbf{x}_p\end{bmatrix} = \begin{bmatrix}\mathbf{X}_{\tilde{p}}^\intercal\mathbf{X}_{\tilde{p}} & \mathbf{X}_{\tilde{p}}^\intercal\mathbf{x}_p \\ \mathbf{x}_p^\intercal\mathbf{X}_{\tilde{p}} & \mathbf{x}_p^\intercal\mathbf{x}_p\end{bmatrix}.$$

Thus by using Schur complements we obtain

$$(\mathbf{X}^\intercal\mathbf{X})^{-1} = \begin{bmatrix} * & * \\ * & \left(\mathbf{x}_p^\intercal\mathbf{x}_p - \mathbf{x}_p^\intercal\mathbf{X}_{\tilde{p}}(\mathbf{X}_{\tilde{p}}^\intercal\mathbf{X}_{\tilde{p}})^{-1}\mathbf{X}_{\tilde{p}}^\intercal\mathbf{x}_p\right)^{-1} \end{bmatrix}$$
$$= \begin{bmatrix} * & * \\ * & \left(\mathbf{x}_p^\intercal\mathbf{x}_p - \mathbf{x}_p^\intercal\mathbf{H}_{\tilde{p}}\mathbf{x}_p\right)^{-1} \end{bmatrix}.$$

Hence we have

$$\begin{aligned} \mathrm{SE}(\hat{\beta}_p)^2 &= \sigma_\epsilon^2\left(\mathbf{x}_p^\intercal\mathbf{x}_p - \mathbf{x}_p^\intercal\mathbf{H}_{\tilde{p}}\mathbf{x}_p\right)^{-1} \\ &= \sigma_\epsilon^2\left(\mathbf{x}_p^\intercal\mathbf{x}_p - \mathbf{x}_p^\intercal\tilde{\mathbf{x}}_p\right)^{-1} = \sigma_\epsilon^2\left(\mathbf{x}_p^\intercal(\mathbf{x}_p - \tilde{\mathbf{x}}_p)\right)^{-1} \\ &= \sigma_\epsilon^2\left(\|\mathbf{x}_p - \tilde{\mathbf{x}}_p\|^2 + \tilde{\mathbf{x}}_p^\intercal(\mathbf{x}_p - \tilde{\mathbf{x}}_p)\right)^{-1} = \frac{\sigma_\epsilon^2}{\|\mathbf{x}_p - \tilde{\mathbf{x}}_p\|^2}, \end{aligned}$$

since $\mathbf{x}_p - \tilde{\mathbf{x}}_p$ is orthogonal to V_{p-1}, which contains $\tilde{\mathbf{x}}_p$. (Note that this provides a new formula, and interpretation, for $\mathrm{SE}(\hat{\beta}_p)^2$.)

Finally we get

$$t^2 = \frac{\hat{\beta}_p^2}{\widehat{\mathrm{SE}}(\hat{\beta}_p)^2} = \frac{\hat{\beta}_p^2\|\mathbf{x}_p - \tilde{\mathbf{x}}_p\|^2}{\hat{\sigma}_\epsilon^2} = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/1}{\mathrm{RSS}/(n-p-1)} = F,$$

as desired. The $\hat{\sigma}_\epsilon^2$ is an estimate of σ_ϵ^2, which here we substituted RSE for it.

The problem with individual t-statistics is that when there are many predictors, some of their associated p-values can be small due to chance. And these randomly small p-values do not imply that there is a relationship between the response and their corresponding predictors. Therefore we use the F-statistic to see if there is a relation between the response and predictors in multiple regression. Notice that in the joint distribution of all the t-statistics, the event of some of their p-values being small is quite likely! So the smallness of some of the p-values is not an improbable event, and based on it we cannot reject the null hypothesis. In contrast, the smallness of the p-value of the F-statistic is an improbable event, and based on it we can reject H_0.

2.2.6 R^2 Statistic in Multiple Linear Regression

In multiple linear regression R^2 is equal to $\mathrm{Cor}^2(Y, \hat{Y})$, i.e. the square of the correlation between the response and its predicted value in the linear model. In addition, the least squares linear model maximizes this correlation (not its square) among all possible linear models. Let us prove these claims. First note that $\mathbf{y} - \hat{\mathbf{y}}$ is

orthogonal to the image of $\cdot \mapsto \mathbf{X}\cdot$, so it is in particular orthogonal to

$$\mathbf{X}\begin{bmatrix}\hat{\beta}_0 - \bar{y}\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_p\end{bmatrix} = \begin{bmatrix}1 & x_{11} & x_{12} & \dots & x_{1p}\\ 1 & x_{21} & x_{22} & \dots & x_{2p}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{n1} & x_{n2} & \dots & x_{np}\end{bmatrix}\begin{bmatrix}\hat{\beta}_0 - \bar{y}\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_p\end{bmatrix} = \begin{bmatrix}\hat{y}_1 - \bar{y}\\ \hat{y}_2 - \bar{y}\\ \vdots\\ \hat{y}_n - \bar{y}\end{bmatrix}.$$

Hence the inner product of $\mathbf{y} - \hat{\mathbf{y}}$ and the above vector is zero, i.e.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0.$$

Thus we have

$$\begin{aligned}\mathrm{TSS} - \mathrm{RSS} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2\\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})\\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)(y_i - \hat{y}_i + \hat{y}_i - \bar{y})\\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)(y_i - \bar{y})\\ &= \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y} - y_i + \hat{y}_i) = \sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y}).\end{aligned}$$

Now suppose $\mathbf{z}$ is the prediction of Y using some other linear model. Then we have $\mathbf{z} = \mathbf{X}\boldsymbol{\gamma}$ for some vector $\boldsymbol{\gamma}$. Similarly to the above, we can see that $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to $\begin{bmatrix}z_1 - \bar{z} & \dots & z_n - \bar{z}\end{bmatrix}^\intercal$. Hence their inner product is zero too, i.e.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(z_i - \bar{z}) = 0.$$

Therefore we have

$$\begin{aligned}(n-1)q_{yz} &= \sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z})\\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})(z_i - \bar{z})\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}(y_i - \hat{y}_i)(z_i - \bar{z}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})(z_i - \bar{z}) \\
&= 0 + \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(z_i - \bar{z}) \qquad \text{(since } \bar{\hat{y}} = \bar{y}\text{)} \\
&= (n-1)q_{\hat{y}z}.
\end{aligned}
$$

So $q_{yz} = q_{\hat{y}z}$. In particular for $\mathbf{z} = \hat{\mathbf{y}}$ we have

$$q_{y\hat{y}} = q_{\hat{y}\hat{y}} = s^2_{\hat{y}}.$$

Hence we get

$$
\begin{aligned}
R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\
&= \frac{q_{y\hat{y}}}{s^2_y} = \frac{q_{y\hat{y}}}{s^2_y}\frac{s^2_{\hat{y}}}{s^2_{\hat{y}}} = \frac{q_{y\hat{y}}}{s^2_y}\frac{q_{y\hat{y}}}{s^2_{\hat{y}}} = \frac{q^2_{y\hat{y}}}{s^2_y s^2_{\hat{y}}} = \text{Cor}^2(Y, \hat{Y}),
\end{aligned}
$$

as desired.

On the other hand, we have

$$
\begin{aligned}
\text{Cor}(Y, Z) = \frac{q_{yz}}{s_y s_z} &= \frac{q_{\hat{y}z}}{s_y s_z} = \frac{s^2_{\hat{y}}}{s^2_{\hat{y}}}\frac{q_{\hat{y}z}}{s_y s_z} = \frac{s^2_{\hat{y}}}{s_y s_{\hat{y}}}\frac{q_{\hat{y}z}}{s_{\hat{y}} s_z} = \frac{q_{\hat{y}y}}{s_y s_{\hat{y}}}\frac{q_{\hat{y}z}}{s_{\hat{y}} s_z} \\
&= \text{Cor}(Y, \hat{Y})\text{Cor}(\hat{Y}, Z) \leq \text{Cor}(Y, \hat{Y}) \cdot 1 = \text{Cor}(Y, \hat{Y}).
\end{aligned}
$$

Note that in the above inequality we used the facts that $\text{Cor}(\hat{Y}, Z) \leq 1$, and

$$\text{Cor}(Y, \hat{Y}) = \frac{q_{y\hat{y}}}{s_y s_{\hat{y}}} = \frac{s^2_{\hat{y}}}{s_y s_{\hat{y}}} = \frac{s_{\hat{y}}}{s_y} \geq 0.$$

As a side note, observe that if we use the inequality $\text{Cor}(Y, \hat{Y}) \leq 1$ in the above equation we obtain

$$s_{\hat{y}} \leq s_y.$$

In other words, we have also shown that the correlation of $Y, \hat{Y}$ cannot be negative, and the variance of $\hat{Y}$ cannot be larger than the variance of Y.

2.2.7 Confidence and Prediction Intervals

Remember that we have

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon}.$$

This relation allowed us to compute the expected value and the covariance matrix of $\hat{\boldsymbol{\beta}}$, yielding the formulas

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}, \qquad \mathbb{E}\big[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\intercal}\big] = \sigma_{\epsilon}^2 (\mathbf{X}^{\intercal}\mathbf{X})^{-1}.$$

Now suppose that based on $\hat{\boldsymbol{\beta}}$ we make a prediction at a new observation of predictors

$$\boldsymbol{x}_0^{\intercal} = \begin{bmatrix} 1 & x_{01} & x_{02} & \dots & x_{0p} \end{bmatrix},$$

to obtain the estimate $\hat{y}_0 = \boldsymbol{x}_0^{\intercal}\hat{\boldsymbol{\beta}}$ for the response value $y_0 = \boldsymbol{x}_0^{\intercal}\boldsymbol{\beta} + \epsilon_0$. Then the expected value and variance of $\hat{y}_0$ are

$$\begin{aligned}
\mathbb{E}[\hat{y}_0] &= \mathbb{E}[\boldsymbol{x}_0^{\intercal}\hat{\boldsymbol{\beta}}] = \boldsymbol{x}_0^{\intercal}\,\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{x}_0^{\intercal}\boldsymbol{\beta}, \\
\mathrm{Var}[\hat{y}_0] &= \mathbb{E}[(\hat{y}_0 - \mathbb{E}[\hat{y}_0])^2] = \mathbb{E}[(\boldsymbol{x}_0^{\intercal}\hat{\boldsymbol{\beta}} - \boldsymbol{x}_0^{\intercal}\boldsymbol{\beta})^2] \\
&= \mathbb{E}[(\boldsymbol{x}_0^{\intercal}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))(\boldsymbol{x}_0^{\intercal}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))] \\
&= \mathbb{E}[(\boldsymbol{x}_0^{\intercal}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))(\boldsymbol{x}_0^{\intercal}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^{\intercal}] \qquad \text{(since } a^{\intercal} = a \text{ for a scalar)} \\
&= \mathbb{E}[\boldsymbol{x}_0^{\intercal}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\intercal}\boldsymbol{x}_0] \\
&= \boldsymbol{x}_0^{\intercal}\,\mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\intercal}]\boldsymbol{x}_0 = \sigma_{\epsilon}^2\,\boldsymbol{x}_0^{\intercal}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\boldsymbol{x}_0.
\end{aligned}$$

Therefore the expected value and variance of $\boldsymbol{x}_0^{\intercal}\boldsymbol{\beta} - \hat{y}_0$ are

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_0^{\intercal}\boldsymbol{\beta} - \hat{y}_0] &= \mathbb{E}[\boldsymbol{x}_0^{\intercal}\boldsymbol{\beta}] - \mathbb{E}[\hat{y}_0] = 0, \qquad \text{(since } \boldsymbol{\beta} \text{ is constant)} \\
\mathrm{Var}[\boldsymbol{x}_0^{\intercal}\boldsymbol{\beta} - \hat{y}_0] &= \mathbb{E}[(\boldsymbol{x}_0^{\intercal}\boldsymbol{\beta} - \hat{y}_0)^2] \\
&= \mathbb{E}[(\hat{y}_0 - \mathbb{E}[\hat{y}_0])^2] = \mathrm{Var}[\hat{y}_0] = \sigma_{\epsilon}^2\,\boldsymbol{x}_0^{\intercal}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\boldsymbol{x}_0.
\end{aligned}$$

This variance can be used to compute **confidence intervals**, which are estimates of how much $\boldsymbol{x}_0^{\intercal}\boldsymbol{\beta}$ (which is the expected value of y_0 given $\boldsymbol{x}_0$, since $\mathbb{E}[\epsilon_0] = 0$) differs from the predicted value $\hat{y}_0$.

Next let us find the expected value and variance of $y_0 - \hat{y}_0$. We have

$$\mathbb{E}[y_0 - \hat{y}_0] = \mathbb{E}[\boldsymbol{x}_0^{\intercal}\boldsymbol{\beta} + \epsilon_0 - \hat{y}_0] = \boldsymbol{x}_0^{\intercal}\boldsymbol{\beta} + \mathbb{E}[\epsilon_0] - \mathbb{E}[\hat{y}_0] = \boldsymbol{x}_0^{\intercal}\boldsymbol{\beta} + 0 - \boldsymbol{x}_0^{\intercal}\boldsymbol{\beta} = 0.$$

Note that $\epsilon_0, \hat{\boldsymbol{\beta}}$ are independent, because $\epsilon_0, \boldsymbol{\epsilon}$ are independent. Hence we get

$$\begin{aligned}
\mathrm{Var}[y_0 - \hat{y}_0] &= \mathbb{E}[(y_0 - \hat{y}_0)^2] = \mathbb{E}[(\boldsymbol{x}_0^{\intercal}\boldsymbol{\beta} + \epsilon_0 - \boldsymbol{x}_0^{\intercal}\hat{\boldsymbol{\beta}})^2] \\
&= \mathbb{E}[(\boldsymbol{x}_0^{\intercal}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^2] - 2\mathbb{E}[\boldsymbol{x}_0^{\intercal}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\epsilon_0] + \mathbb{E}[\epsilon_0^2] \\
&= \sigma_{\epsilon}^2\,\boldsymbol{x}_0^{\intercal}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\boldsymbol{x}_0 - 2\boldsymbol{x}_0^{\intercal}\,\mathbb{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}]\mathbb{E}[\epsilon_0] + \mathrm{Var}[\epsilon_0] \\
&= \sigma_{\epsilon}^2\,\boldsymbol{x}_0^{\intercal}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\boldsymbol{x}_0 - 0 + \sigma_{\epsilon}^2 = \sigma_{\epsilon}^2\big(\boldsymbol{x}_0^{\intercal}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\boldsymbol{x}_0 + 1\big).
\end{aligned}$$

The above variance can be used to compute **prediction intervals**, which are estimates of how much y_0 differs from the predicted value $\hat{y}_0$.

Note that prediction intervals provide upper and lower estimates for a single value of response, i.e. y_0. In contrast, confidence intervals provide such estimates for the average value of response. The reason is that the expected value of ϵ_0 is zero, thus its effects are expected to diminish in average, leaving us with $\boldsymbol{x}_0^\intercal \boldsymbol{\beta}$ as the average value of response.

Also note that both prediction and confidence intervals are centered around the predicted value $\hat{y}_0$, but the prediction interval is larger, since the variance of $y_0 - \hat{y}_0$ is larger than the variance of $\boldsymbol{x}_0^\intercal \boldsymbol{\beta} - \hat{y}_0$. Another way of looking at this is that the variance of $\boldsymbol{x}_0^\intercal \boldsymbol{\beta} - \hat{y}_0$ consists of the reducible error, while the variance of $y_0 - \hat{y}_0$ consists of both reducible and irreducible errors.

2.3 Further Topics in Linear Regression

2.3.1 Polynomial Regression and Basis Functions

We can use the method of least squares to fit nonlinear functions instead of linear functions. For example, we can fit polynomial functions to the data. To this end, we still use the linear regression approach, but we use new features which are themselves nonlinear functions of the original features. For example, instead of just using a feature X, we can use the features

$$X_1 = X, \quad X_2 = X^2, \quad \dots \quad X_m = X^m,$$

to find a degree m polynomial that minimizes the sum of squared residuals. In other words, we can estimate the coefficients in a relation of the form

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \epsilon \\ &= \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_m X^m + \epsilon, \end{aligned}$$

by the usual method of least squares in multiple linear regression. This is known as **polynomial regression**.

When we initially have more than one features, say $X_1, \dots, X_p$, we can use new features which are monomial functions of several variables. In particular, we can also use **interaction terms** of the form $X_1 X_2$ as new features.

We can use the above idea to fit functions other than polynomials too. In general, we can use the new features

$$b_1(X), b_2(X), \dots, b_m(X)$$

in a multiple linear regression, where $b_1, \dots, b_m$ are given functions. (Note that X can itself be a vector of several features, in which case $b_1, \dots, b_m$ will be functions of several variables.) The functions $b_1, \dots, b_m$ are called **basis functions**.

2.3.2 Qualitative Predictors and Dummy Variables

Let X be a quantitative feature whose range contains the values $c_1, \ldots, c_k$. A particular choice of basis functions are piecewise constant functions:

$$\begin{aligned} C_0(X) &= I(X < c_1), \\ C_1(X) &= I(c_1 \le X < c_2), \\ &\vdots \\ C_{k-1}(X) &= I(c_{k-1} \le X < c_k), \\ C_k(X) &= I(X \ge c_k), \end{aligned}$$

where I is the **indicator function** whose value is 1 if the inside condition is satisfied, and 0 otherwise. The new features $C_0(X), \ldots, C_k(X)$ are also called **dummy variables**. Note that a linear combination of $C_0(X), \ldots, C_k(X)$ is a piecewise constant function, i.e. it is constant on each of the intervals $[c_j, c_{j+1})$. Note that we always have

$$C_0(X) + C_1(X) + \cdots + C_k(X) = 1,$$

since X belongs to exactly one the intervals. Hence there is a linear dependence relation between the dummy variables and the constant function 1. So if we use the intercept term in the regression, which is itself a linear combination of the constant function 1, we have to put aside one of the dummy variables, and work with the rest of them. The choice of the dropped dummy variable is arbitrary due to the above linear dependence relation.

The idea of dummy variables is particularly helpful when X is a qualitative feature with k levels $c_1, \ldots, c_k$ (note that $c_1, \ldots, c_k$ are not necessarily numbers here). In this case the dummy variables become

$$\begin{aligned} C_1(X) &= I(X = c_1), \\ &\vdots \\ C_{k-1}(X) &= I(X = c_{k-1}), \\ C_k(X) &= I(X = c_k). \end{aligned}$$

These new variables are quantitative variables, and can be used in an ordinary least squares regression. So we can deal with the case where some of the features are qualitative variables.

Note that each dummy variable has only two levels, and we use k two level dummy variables to quantify the qualitative variable X, and we do not use a single k-level dummy variable. The reason is that if we want to use a dummy variable with

more than two levels, we have to code the levels of X by numbers. And in doing so we will impose a linear order on the levels of X, together with a notion of distance between those levels. However, in general, the levels of a qualitative variable are not ordered linearly, nor there is a well-defined notion of distance between them. Hence we use several two-level dummy variables instead. In this way, we avoid imposing any order on the levels. Also, we do not assign any notion of distance between different levels, since the 0/1 coding of belonging or not belonging to a class merely expresses that classes are different.

2.3.3 Heteroscedasticity and Weighted Least Squares

So far we have assumed that in the relation $Y = f(X)+\epsilon$, the error term ϵ does not depend on X. Note that the variance of Y under the condition that X is constant is

$$\mathrm{Var}(Y) = \mathrm{Var}(f(X) + \epsilon) = \mathrm{Var}(\epsilon).$$

Hence if ϵ does not depend on X then $\mathrm{Var}(Y)$ does not depend on the value of X either. However, in general, this assumption does not hold, and the variance of Y can depend on the value of X. This phenomenon is known as **heteroscedasticity**. In the presence of heteroscedasticity (and under the assumption that f is linear), the observed data will satisfy

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

in which ϵ_is are independent random variables with zero mean, whose variances σ_i^2 depend on i, unlike before.

To estimate the coefficients $\beta_0, \ldots, \beta_p$ under the above assumptions we can use the **weighted least squares** method. In this method, instead of minimizing RSS, we minimize a weighted version of it:

$$\sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} w_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2,$$

where $w_i = \frac{1}{\sigma_i^2}$. Let $\mathbf{W}$ be the diagonal matrix with diagonal entries w_i. Then the above weighted sum of squares can be written as

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\intercal \mathbf{W} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

By differentiating this expression with respect to $\hat{\boldsymbol{\beta}}$, similarly to Section 2.2.1, we can easily show that the minimizer is given by

$$\mathbf{X}^\intercal \mathbf{W} \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^\intercal \mathbf{W} \mathbf{y},$$

or equivalently by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{W} \mathbf{y},$$

when $\mathbf{X}^\intercal \mathbf{W} \mathbf{X}$ is invertible.

2.3.4 Outliers and Studentized residuals

Outliers are data points which are far from the rest of the data. At an outlier the response has an unusual value. The predictors may have unusual values at a point too. In this section and the next one we consider statistics that can help us detect outliers and points with outlying predictor values.

Recall that $\mathbf{H} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$ is the **projection (hat) matrix** on the image of the linear map defined by $\mathbf{X}$. Then we have $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. Now note that

$$\mathbf{H}\mathbf{X} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}(\mathbf{X}^\intercal\mathbf{X}) = \mathbf{X}.$$

Therefore $(I - \mathbf{H})\mathbf{X} = \mathbf{X} - \mathbf{H}\mathbf{X} = \mathbf{X} - \mathbf{X} = 0$. Hence we get

$$\mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (I - \mathbf{H})\mathbf{y} = (I - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (I - \mathbf{H})\boldsymbol{\epsilon}.$$

In addition, note that we have

$$(I - \mathbf{H})^2 = I - I\mathbf{H} - \mathbf{H}I + \mathbf{H}^2 = I - 2\mathbf{H} + \mathbf{H} = I - \mathbf{H},$$

since $\mathbf{H}^2 = \mathbf{H}$ (because $\mathbf{H}$ is a projection matrix). In fact, it can be shown that $I - \mathbf{H}$ is the projection matrix on the orthogonal complement of the image of $\mathbf{X}$; so its square must be equal to itself. This also implies that $(I - \mathbf{H})\mathbf{X}$ must be zero, because the columns of $\mathbf{X}$ belong to its image, and the projection of the image of $\mathbf{X}$ onto its orthogonal complement is zero.

Now let us compute the covariance matrix of $\mathbf{y} - \hat{\mathbf{y}}$. First note that

$$\mathbb{E}[\mathbf{y} - \hat{\mathbf{y}}] = \mathbb{E}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} - \mathbf{X}\hat{\boldsymbol{\beta}}] = \mathbf{X}\boldsymbol{\beta} + 0 - \mathbf{X}\,\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = 0.$$

Thus we have

$$\begin{aligned}
\mathbb{E}\big[(\mathbf{y} - \hat{\mathbf{y}})(\mathbf{y} - \hat{\mathbf{y}})^\intercal\big] &= \mathbb{E}\big[(I - \mathbf{H})\boldsymbol{\epsilon}((I - \mathbf{H})\boldsymbol{\epsilon})^\intercal\big] \\
&= \mathbb{E}\big[(I - \mathbf{H})\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal(I - \mathbf{H})^\intercal\big] \\
&= (I - \mathbf{H})\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal](I - \mathbf{H})^\intercal && \text{(since } \mathbf{H} \text{ is constant)} \\
&= (I - \mathbf{H})\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal](I - \mathbf{H}) && \text{(since } \mathbf{H} \text{ is symmetric)} \\
&= (I - \mathbf{H})\sigma_\epsilon^2 I(I - \mathbf{H}) = \sigma_\epsilon^2(I - \mathbf{H})^2 = \sigma_\epsilon^2(I - \mathbf{H}).
\end{aligned}$$

Therefore we have $\mathrm{Var}(e_i) = \mathrm{Var}(y_i - \hat{y}_i) = (\sigma_\epsilon^2(I - \mathbf{H}))_{ii} = \sigma_\epsilon^2(1 - h_i)$, where h_i is the ith diagonal entry of $\mathbf{H}$, and is known as the leverage (discussed below). This enables us to compute the standard error of the residual e_i:

$$\mathrm{SE}(e_i) = \sigma_\epsilon\sqrt{1 - h_i}.$$

The **studentized residual** t_i is defined as the ratio of the residual e_i to its standard error, i.e.

$$t_i = \frac{e_i}{\sigma_\epsilon\sqrt{1 - h_i}}.$$

Note that when the leverage h_i is close to 1 (which is its maximum as we will see below), or the residual e_i is large, then the studentized residual t_i will be large. An observation with a large studentized residual can be an outlier.

Also note that the variance of $e_i = y_i - \hat{y}_i$ cannot be computed similarly to the variance of $y_0 - \hat{y}_0$ (which is computed in the previous subsection). Because unlike ϵ_0, the error term ϵ_i is not independent from the random vector $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_i, \ldots, \epsilon_n]^\intercal$.

2.3.5 Leverage

The **leverage statistic** h_i is the ith diagonal entry of the projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$. By using the formula for $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ we obtain

$$\mathbf{H} = \frac{1}{n-1}\mathbf{X}\begin{bmatrix} n^{-1}(n-1) + \bar{\boldsymbol{x}}^\intercal Q^{-1}\bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}^\intercal Q^{-1} \\ -Q^{-1}\bar{\boldsymbol{x}} & Q^{-1} \end{bmatrix}\mathbf{X}^\intercal,$$

where Q is the matrix of sample covariances, and $\bar{\boldsymbol{x}}$ is the vector of sample means. Let $\boldsymbol{x}_i = \begin{bmatrix} x_{i1} & \ldots & x_{ip} \end{bmatrix}^\intercal$ be the vector of ith observation. Then the ith row of $\mathbf{X}$ is $\begin{bmatrix} 1 & \boldsymbol{x}_i^\intercal \end{bmatrix}$. Hence we get

$$\begin{aligned} h_i &= \frac{1}{n-1}\begin{bmatrix} 1 & \boldsymbol{x}_i^\intercal \end{bmatrix}\begin{bmatrix} n^{-1}(n-1) + \bar{\boldsymbol{x}}^\intercal Q^{-1}\bar{\boldsymbol{x}} & -\bar{\boldsymbol{x}}^\intercal Q^{-1} \\ -Q^{-1}\bar{\boldsymbol{x}} & Q^{-1} \end{bmatrix}\begin{bmatrix} 1 \\ \boldsymbol{x}_i \end{bmatrix} \\ &= \frac{1}{n-1}\begin{bmatrix} 1 & \boldsymbol{x}_i^\intercal \end{bmatrix}\begin{bmatrix} n^{-1}(n-1) + \bar{\boldsymbol{x}}^\intercal Q^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{x}_i) \\ -Q^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{x}_i) \end{bmatrix} \\ &= \frac{1}{n} + \frac{1}{n-1}(\bar{\boldsymbol{x}} - \boldsymbol{x}_i)^\intercal Q^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{x}_i). \end{aligned}$$

Now note that Q is positive definite since as we have seen before, Q is invertible, and satisfies

$$(n-1)Q = (\tilde{\mathbf{X}} - \bar{\boldsymbol{x}}\mathbf{1}^\intercal)^\intercal(\tilde{\mathbf{X}} - \bar{\boldsymbol{x}}\mathbf{1}^\intercal),$$

where $\mathbf{X} = \begin{bmatrix} \mathbf{1} & \tilde{\mathbf{X}} \end{bmatrix}$. Therefore Q^{-1} is also positive definite. Thus the nonnegative number

$$(\bar{\boldsymbol{x}} - \boldsymbol{x}_i)^\intercal Q^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{x}_i)$$

measures the square of the distance of the ith observation from the mean, with respect to the inner product induced by Q^{-1}. This distance is known as the **Mahalanobis distance**, and intuitively, it measures the distance between a point $\boldsymbol{x}_i$ and the mean $\bar{\boldsymbol{x}}$ in the units of standard deviations. Hence an observation with high leverage has outlying predictor values. In the simple linear regression the covariance matrix Q has only one entry, i.e. the sample variance; hence the leverage statistic is given by

$$h_i = \frac{1}{n} + \frac{1}{n-1}\frac{(x_i - \bar{x})^2}{s_x^2} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}.$$

Another property of h_i, which motivates the name leverage, is that

$$h_i = \frac{\partial \hat{y}_i}{\partial y_i},$$

since we have $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. So at points with high leverage, small errors in measurement of y can result in significant errors in the predicted value $\hat{y}$. This can also be seen from the fact that the variance of $\hat{y}_i$ is a multiple of h_i. To show this remember that $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Hence

$$\mathbb{E}[\hat{\mathbf{y}}] = \mathbb{E}[\mathbf{X}\hat{\boldsymbol{\beta}}] = \mathbf{X}\,\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbf{X}\boldsymbol{\beta}.$$

Therefore the covariance matrix of $\hat{\mathbf{y}}$ is

$$\begin{aligned}
\mathbb{E}\big[(\hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})(\hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^\intercal\big] &= \mathbb{E}\big[(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})^\intercal\big] \\
&= \mathbb{E}\big[\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^\intercal\big] \\
&= \mathbb{E}\big[\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\intercal\mathbf{X}^\intercal\big] \\
&= \mathbf{X}\,\mathbb{E}\big[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\intercal\big]\mathbf{X}^\intercal \qquad \text{(since } \mathbf{X} \text{ is constant)} \\
&= \mathbf{X}\,\sigma_\epsilon^2(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal = \sigma_\epsilon^2\,\mathbf{H}.
\end{aligned}$$

So we get $\mathrm{Var}(\hat{y}_i) = (\sigma_\epsilon^2\,\mathbf{H})_{ii} = \sigma_\epsilon^2 h_i$. Notice that this variance cannot be computed similarly to the case of $\hat{y}_0$ (which is computed in a previous subsection), because as we noted before, unlike ϵ_0, the error term ϵ_i is not independent from the random vector $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_i, \ldots, \epsilon_n]^\intercal$.

The average leverage for all the observations is always equal to $(p+1)/n$, because

$$\sum h_i = \mathrm{tr}(\mathbf{H}) = \mathrm{tr}\big(\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\big) = \mathrm{tr}\big((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}\big) = \mathrm{tr}(I) = p + 1.$$

Furthermore, the leverage statistic is always between $1/n$ and 1. We have already seen that $h_i \geq 1/n$. To show the other bound note that we have $\mathbf{H}^\intercal = \mathbf{H} = \mathbf{H}^2$, since $\mathbf{H}$ is the matrix of an orthogonal projection. Hence

$$h_i = \mathbf{H}_{ii} = (\mathbf{H}^2)_{ii} = \sum_j \mathbf{H}_{ij}\mathbf{H}_{ji} = \sum_j (\mathbf{H}_{ij})^2 = h_i^2 + \sum_{j \neq i} (\mathbf{H}_{ij})^2;$$

so $h_i^2 \leq h_i$, which implies $h_i \leq 1$.

2.3.6 Collinearity

Suppose that we can estimate the predictor X_j using the other predictors quite accurately. This phenomenon is known as **(multi)collinearity**. When collinearity is present, our uncertainty in our estimate of the corresponding regression coefficient will be high, since we know that

$$\mathrm{SE}(\hat{\beta}_j)^2 = \frac{\sigma_\epsilon^2}{\|\mathbf{x}_j - \tilde{\mathbf{x}}_j\|^2}, \qquad (*)$$

where $\tilde{\mathbf{x}}_j$ is the prediction of $\mathbf{x}_j$ based on the other predictors. As a result, the corresponding t-statistic will be small, and we may fail to reject the null hypothesis $H_0 : \beta_j = 0$.

To assess the severity of multicollinearity we can use the **variance inflation factor (VIF)**, which is the ratio of the variance of $\hat{\beta}_j$ in the full model with all the predictors to its variance in the model with only one predictor X_j, i.e.

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{\mathrm{SE}(\hat{\beta}_j)^2}{\mathrm{SE}(\hat{\beta}_{j,\text{ simple}})^2}.$$

In other words, VIF measures the change in our uncertainty in $\hat{\beta}_j$ when we extend the model with only one predictor X_j by adding the other predictors.

Let $R^2_{X_j|X_{-j}}$ be the R^2 statistic from regressing X_j onto all the other predictors. Let $\mathrm{RSS}_{j|-j}, \mathrm{TSS}_{j|-j}$ be respectively the RSS and TSS of this regression. Also let $\tilde{\mathbf{x}}_j = \begin{bmatrix}\tilde{x}_{1j} & \dots & \tilde{x}_{nj}\end{bmatrix}^\intercal$ be the prediction of $\mathbf{x}_j = \begin{bmatrix}x_{1j} & \dots & x_{nj}\end{bmatrix}^\intercal$ based on the other predictors. Then we have

$$\begin{aligned}
\frac{1}{1 - R^2_{X_j|X_{-j}}} &= \frac{1}{\frac{\mathrm{RSS}_{j|-j}}{\mathrm{TSS}_{j|-j}}} = \frac{\mathrm{TSS}_{j|-j}}{\mathrm{RSS}_{j|-j}} \\
&= \frac{\sum_i (x_{ij} - \bar{x}_j)^2}{\sum_i (x_{ij} - \tilde{x}_{ij})^2} = \frac{\sum_i (x_{ij} - \bar{x}_j)^2}{\|\mathbf{x}_j - \tilde{\mathbf{x}}_j\|^2} \\
&= \frac{\sigma_\epsilon^2 / \|\mathbf{x}_j - \tilde{\mathbf{x}}_j\|^2}{\sigma_\epsilon^2 / \sum_i (x_{ij} - \bar{x}_j)^2} = \frac{\mathrm{SE}(\hat{\beta}_j)^2}{\mathrm{SE}(\hat{\beta}_{j,\text{ simple}})^2} = \mathrm{VIF}(\hat{\beta}_j).
\end{aligned}$$

As a result we can see that the smallest possible value for VIF is 1, since $R^2_{X_j|X_{-j}}$ is between zero and 1 (remember that $R^2_{X_j|X_{-j}}$ is equal to the square of correlation of X_j and its prediction $\tilde{X}_j$). When VIF is 1 we must have $R^2_{X_j|X_{-j}} = 0$; thus the other predictors provide no information about X_j, and there is no collinearity. On the other hand, if $R^2_{X_j|X_{-j}}$ is close to 1 then there is considerable collinearity, and VIF has a large value.

We can also see that the smallest possible value for VIF is 1 by noting that $\mathrm{SE}(\hat{\beta}_j)^2$ increases as we add more predictors to the model. The reason is that our prediction of $\mathbf{x}_j$ based on the other predictors becomes more accurate when we add more predictors to the model, because our prediction of $\mathbf{x}_j$ is the orthogonal projection of $\mathbf{x}_j$ onto the subspace spanned by our observations of the other predictors. Hence as we add more predictors to the model, this subspace becomes larger. And therefore the projection of $\mathbf{x}_j$ becomes closer to $\mathbf{x}_j$, since we know that the orthogonal projection of a vector like $\mathbf{x}_j$ onto a subspace is the closest vector in that subspace to the given vector $\mathbf{x}_j$; and if we enlarge the subspace the new

closest vector to $\mathbf{x}_j$ cannot be farther away from $\mathbf{x}_j$ than the old closest vector. Therefore by formula $(*)$ the $\mathrm{SE}(\hat{\beta}_j)^2$ increases as we add more predictors to the model. (Note that in the case of simple linear regression, the prediction of $\mathbf{x}_j$ using no other predictor is $\bar{x}_j\mathbf{1}$; and plugging this value in the formula $(*)$ gives us $\hat{\beta}_{j,\text{ simple}}$.)

2.3.7 Maximum Likelihood Estimation

Suppose the probability measure $\mathbb{P}$ describes the distribution of some quantities among a population. We assume that $\mathbb{P}$ belongs to a parametric family of probability measures $\mathbb{P}_\theta$, but we do not know which value of θ gives us $\mathbb{P}$. In other words, $\mathbb{P} = \mathbb{P}_{\theta_0}$ for some unknown θ_0. Let $f(x;\theta)$ denote the probability density (mass) function of $\mathbb{P}_\theta$. Suppose $X_1, \ldots, X_n$ are a random sample from this distribution, and $x_1, \ldots, x_n$ are their observed values. (Note that X_is can be random vectors, and so x_is can be vectors too.) Then, the **likelihood function** of θ is the joint probability density (mass) function of $X_1, \ldots, X_n$ at $x_1, \ldots, x_n$, which, if $X_1, \ldots, X_n$ are independent, is equal to

$$L(\theta) = L(\theta | x_1, \ldots, x_n) = \prod_{i \leq n} f(x_i; \theta).$$

Note that the likelihood function is a function of θ, and $x_1, \ldots, x_n$ are considered as (fixed) parameters here.

Now remember that we do not know the value of θ_0, and we would like to estimate it by using the training data $x_1, \ldots, x_n$. The **maximum likelihood** estimate for θ_0 is a value $\hat{\theta}_0$ such that $L(\hat{\theta}_0) \geq L(\theta)$ for all θ, i.e. a point of absolute maximum of L. In other words, $\hat{\theta}_0$ is the value of θ that maximizes the probability of observing $x_1, \ldots, x_n$ (or the probability of observing values close to $x_1, \ldots, x_n$, in the case of continuous probability measures). In practice we usually maximize the **log-likelihood** function

$$\log L(\theta) = \sum_{i \leq n} \log f(x_i; \theta)$$

instead of the likelihood function, since it is easier to work with. Note that the logarithm is a strictly increasing function, so maximizing the log-likelihood results in the same estimate.

Let us see an example of maximum likelihood estimation. Suppose f is normal, i.e.

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2 \Big).$$

(Note that here θ is the vector (μ, σ^2). Also note that here we abuse the notation and represent both θ and θ_0 by (μ, σ^2).) Then the log-likelihood becomes

$$\begin{aligned}\log L(\mu, \sigma^2) &= \sum_{i \le n} \log f(x_i; \mu, \sigma^2) \\ &= \sum_{i \le n} \big(-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i - \mu)^2 \big) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i \le n} (x_i - \mu)^2.\end{aligned}$$

At the point of maximum of $\log L$ we must have

$$0 = \frac{\partial \log L}{\partial \mu}(\hat{\mu}, \hat{\sigma}^2) = \frac{1}{\hat{\sigma}^2} \sum_{i \le n} (x_i - \hat{\mu}) = \frac{1}{\hat{\sigma}^2}(n\bar{x} - n\hat{\mu}).$$

Therefore the maximum likelihood estimate of μ is $\hat{\mu} = \bar{x}$. In other words, the maximum likelihood estimate of the population mean of a normally distributed variable is the sample mean. Similarly we have

$$0 = \frac{\partial \log L}{\partial \sigma^2}(\hat{\mu}, \hat{\sigma}^2) = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i \le n} (x_i - \hat{\mu})^2 = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i \le n} (x_i - \bar{x})^2.$$

(Note that σ^2 is the variable which we differentiated with respect to, not σ!) Hence we have

$$\frac{n}{2\hat{\sigma}^2} = \frac{1}{2\hat{\sigma}^4} \sum_{i \le n} (x_i - \bar{x})^2 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i \le n} (x_i - \bar{x})^2.$$

Chapter 3

Classification

3.1 Logistic Regression

Suppose that the response Y is a qualitative variable with two levels, which we denote by 0 and 1. Usually in classification, instead of the value of Y, we try to estimate the probability

$$p(X) = \mathbb{P}(Y = 1|X).$$

In **logistic regression**, we assume that $p(X)$ is given by the **logistic function**

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}}.$$

It is easy to see that

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}.$$

The ratio $p(X)/[1 - p(X)]$ is called the **odds**. Note that since $p(X)$ is between 0 and 1, the odds is between 0 and ∞. Also note that the odds is a strictly increasing function of the probability $p(X)$. Now by taking logarithm we get

$$\log\big(\frac{p(X)}{1 - p(X)}\big) = \beta_0 + \beta_1 X.$$

The logarithm of the odds is called the **log-odds** or **logit**. Hence in logistic regression we assume that the logit is a linear function of X. Note that logit is also a strictly increasing function of $p(X)$.

To estimate the coefficients β_0, β_1 we use the method of maximum likelihood. In the logistic regression model, we are interested in the conditional probability measure $\mathbb{P}(Y|X)$. (Comparing with the notation of the previous subsection, here θ is the vector (β_0, β_1), and the random vectors $(Y_1, X_1), \ldots, (Y_n, X_n)$ form our

sample.) We want to find the maximum likelihood estimates $\hat{\beta}_0, \hat{\beta}_1$ for β_0, β_1. First note that $\mathbb{P}(Y|X = x_i)$ is a Bernoulli random variable, so we have

$$\begin{aligned} f(y_i, x_i; \theta) &= \mathbb{P}(Y = y_i | X = x_i) \\ &= \begin{cases} p(x_i) & \text{if } y_i = 1 \\ 1 - p(x_i) & \text{if } y_i = 0 \end{cases} = p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}. \end{aligned}$$

Hence the likelihood function is

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_i p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} \\ &= \prod_{y_i=1} p(x_i) \prod_{y_j=0} (1 - p(x_j)) = \prod_{y_i=1} \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \prod_{y_j=0} \frac{1}{1 + e^{\beta_0+\beta_1 x_j}}. \end{aligned}$$

Now we can compute the log-likelihood function and differentiate it to find its point of maximum $(\hat{\beta}_0, \hat{\beta}_1)$. However there is no closed formula for $\hat{\beta}_0, \hat{\beta}_1$, and we have to estimate them using numerical methods.

3.2 Linear Discriminant Analysis

Suppose the response Y is a qualitative variable with levels $\{1, 2, \ldots, K\}$. Let $\pi_k = \mathbb{P}(Y = k)$ be the **prior** probability of $Y = k$, and $f_k(x) = \mathbb{P}(X = x|Y = k)$ be the **density function** of X over $\{Y = k\}$. Then by Bayes' theorem the **posterior** probability of Y given X is equal to

$$\begin{aligned} p_k(x) &= \mathbb{P}(Y = k|X = x) \\ &= \frac{\mathbb{P}(Y = k)\mathbb{P}(X = x|Y = k)}{\mathbb{P}(X = x)} \\ &= \frac{\mathbb{P}(Y = k)\mathbb{P}(X = x|Y = k)}{\sum_{j=1}^K \mathbb{P}(Y = j)\mathbb{P}(X = x|Y = j)} = \frac{\pi_k f_k(x)}{\sum_{j=1}^K \pi_j f_j(x)}. \end{aligned}$$

Note that when X is a continuous random variable we need to work with probability density functions instead of probabilities.

If we further assume that f_k is normal, i.e. $f_k(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\big(-\frac{1}{2\sigma^2}(x - \mu_k)^2\big)$, we get

$$p_k(x) = \frac{\frac{\pi_k}{\sigma\sqrt{2\pi}} \exp\big(-\frac{1}{2\sigma^2}(x - \mu_k)^2\big)}{\sum_{j=1}^K \frac{\pi_j}{\sigma\sqrt{2\pi}} \exp\big(-\frac{1}{2\sigma^2}(x - \mu_j)^2\big)}.$$

To simplify the notation we denote the denominator of $p_k(x)$ by $g(x)$. Then we have

$$\begin{aligned}\log p_k(x) &= \log(\pi_k) - \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}(x-\mu_k)^2 - \log g(x) \\ &= \log(\pi_k) - \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}x^2 + \frac{\mu_k}{\sigma^2}x - \frac{\mu_k^2}{2\sigma^2} - \log g(x) \\ &= \frac{\mu_k}{\sigma^2}x - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) - \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}x^2 - \log g(x).\end{aligned}$$

Hence the value of k that maximizes $p_k(x)$ is the value of k that maximizes

$$\delta_k(x) = \frac{\mu_k}{\sigma^2}x - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k).$$

So this function can be used to determine the Bayes classifier.

Let $x_1, y_1, \ldots, x_n, y_n$ be a set of observations. Let n_k be the number of observations for which $y_i = k$. Consider the estimates

$$\begin{aligned}\hat{\pi}_k &= \frac{n_k}{n}, \\ \hat{\mu}_k &= \frac{1}{n_k}\sum_{y_i=k} x_i, \\ \hat{\sigma}^2 &= \frac{1}{\sum_{k=1}^K (n_k-1)}\sum_{k=1}^K (n_k-1)\Big(\frac{1}{n_k-1}\sum_{y_i=k}(x_i-\hat{\mu}_k)^2\Big) \\ &= \frac{1}{n-K}\sum_{k=1}^K\sum_{y_i=k}(x_i-\hat{\mu}_k)^2.\end{aligned}$$

Note that $\hat{\sigma}^2$ is a weighted average of the sample variances for each of the K classes. Also note that the $n-K$ in the denominator makes $\hat{\sigma}^2$ an unbiased estimator of σ^2, since each $\frac{1}{n_k-1}\sum_{y_i=k}(x_i-\hat{\mu}_k)^2$ is an unbiased estimator of σ^2, and the expectation of a weighted average is equal to the weighted average of the expectations. If we use the above estimates we obtain the following estimate for $\delta_k(x)$:

$$\hat{\delta}_k(x) = \frac{\hat{\mu}_k}{\hat{\sigma}^2}x - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k).$$

This function is called the **discriminant function**. Note that the discriminant function is a linear function in x. Similarly to our use of δ_k in the Bayes classifier, we can use $\hat{\delta}_k$ for classification. This method is known as the **linear discriminant analysis (LDA)**.

Next let us assume that we have more than one predictor, and f_k is multivariate normal, i.e.

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu_k)^\intercal \mathbf{\Sigma}_k^{-1}(x-\mu_k)\Big),$$

where $|\mathbf{\Sigma}_k|$ is the determinant of $\mathbf{\Sigma}_k$. Note that we allow the covariance matrix $\mathbf{\Sigma}_k$ to depend on k. Hence we have

$$p_k(x) = \frac{\pi_k|\mathbf{\Sigma}_k|^{-1/2} \exp\big(-\frac{1}{2}(x-\mu_k)^\intercal \mathbf{\Sigma}_k^{-1}(x-\mu_k)\big)}{\sum_{j=1}^K \pi_j|\mathbf{\Sigma}_j|^{-1/2} \exp\big(-\frac{1}{2}(x-\mu_j)^\intercal \mathbf{\Sigma}_j^{-1}(x-\mu_j)\big)}.$$

As before, let us denote the denominator of $p_k(x)$ by $g(x)$. Then we get

$$\begin{aligned}\log p_k(x) &= \log(\pi_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}(x-\mu_k)^\intercal \mathbf{\Sigma}_k^{-1}(x-\mu_k) - \log g(x) \\ &= -\frac{1}{2}x^\intercal \mathbf{\Sigma}_k^{-1}x + x^\intercal \mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^\intercal \mathbf{\Sigma}_k^{-1}\mu_k \\ &\qquad - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log(\pi_k) - \log g(x).\end{aligned}$$

(Note that we have used the fact that $\mathbf{\Sigma}_k^{-1}$ is symmetric.) Hence the value of k that maximizes $p_k(x)$ is the value of k that maximizes

$$\delta_k(x) = -\frac{1}{2}x^\intercal \mathbf{\Sigma}_k^{-1}x + x^\intercal \mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^\intercal \mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log(\pi_k).$$

So again δ_k can be used to determine the Bayes classifier. We can estimate π_k, μ_k as before, and for the entries of $\mathbf{\Sigma}_k$ we have the estimate

$$(\hat{\mathbf{\Sigma}}_k)_{jl} = \frac{1}{n_k - 1}\sum_{y_i=k}\big((x_i)_j - (\hat{\mu}_k)_j\big)\big((x_i)_l - (\hat{\mu}_k)_l\big),$$

which are simply the sample (co)variances. Therefore we can estimate $\delta_k(x)$ by the discriminant function

$$\hat{\delta}_k(x) = -\frac{1}{2}x^\intercal \hat{\mathbf{\Sigma}}_k^{-1}x + x^\intercal \hat{\mathbf{\Sigma}}_k^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^\intercal \hat{\mathbf{\Sigma}}_k^{-1}\hat{\mu}_k - \frac{1}{2}\log|\hat{\mathbf{\Sigma}}_k| + \log(\hat{\pi}_k).$$

Note that this time the discriminant function is a quadratic function in x. If we use $\hat{\delta}_k$ for classification, as we used δ_k in the Bayes classifier, we obtain the **quadratic discriminant analysis (QDA)** method.

If we assume that the covariance matrix does not depend on k, the above formulas for QDA will reduce to the corresponding formulas for LDA with more that one predictor. Especially, the linear discriminant function is given by

$$\hat{\delta}_k(x) = x^\intercal \hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^\intercal \hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_k + \log(\hat{\pi}_k),$$

where the entries of $\hat{\mathbf{\Sigma}}$ are weighted averages of the sample (co)variances of each of the classes

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{jl} &= \frac{1}{\sum_{k=1}^{K}(n_k-1)}\sum_{k=1}^{K}(n_k-1)\Big(\frac{1}{n_k-1}\sum_{y_i=k}\big((x_i)_j-(\hat{\mu}_k)_j\big)\big((x_i)_l-(\hat{\mu}_k)_l\big)\Big)\\ &= \frac{1}{n-K}\sum_{k=1}^{K}\sum_{y_i=k}\big((x_i)_j-(\hat{\mu}_k)_j\big)\big((x_i)_l-(\hat{\mu}_k)_l\big).\end{aligned}$$

Let us also mention that when Y only has two levels, the log-odds in the LDA framework is linear in x. To see this note that

$$\begin{aligned}\frac{p_1(x)}{1-p_1(x)} &= \frac{\pi_1\exp\big(-\frac{1}{2}(x-\mu_1)^{\intercal}\mathbf{\Sigma}^{-1}(x-\mu_1)\big)}{\pi_2\exp\big(-\frac{1}{2}(x-\mu_2)^{\intercal}\mathbf{\Sigma}^{-1}(x-\mu_2)\big)}\\ &= \frac{\pi_1}{\pi_2}\exp\Big(-\frac{1}{2}\big((x-\mu_1)^{\intercal}\mathbf{\Sigma}^{-1}(x-\mu_1)-(x-\mu_2)^{\intercal}\mathbf{\Sigma}^{-1}(x-\mu_2)\big)\Big)\\ &= \frac{\pi_1}{\pi_2}\exp\Big((\mu_1-\mu_2)^{\intercal}\mathbf{\Sigma}^{-1}x-\frac{1}{2}\mu_1^{\intercal}\mathbf{\Sigma}^{-1}\mu_1+\frac{1}{2}\mu_2^{\intercal}\mathbf{\Sigma}^{-1}\mu_2\Big).\end{aligned}$$

Therefore

$$\log\Big(\frac{p_1(x)}{1-p_1(x)}\Big) = \log\big(\frac{\pi_1}{\pi_2}\big)-\frac{1}{2}\mu_1^{\intercal}\mathbf{\Sigma}^{-1}\mu_1+\frac{1}{2}\mu_2^{\intercal}\mathbf{\Sigma}^{-1}\mu_2+(\mu_1-\mu_2)^{\intercal}\mathbf{\Sigma}^{-1}x,$$

as desired.

Chapter 4

Model Selection and Regularization

4.1 Cross-Validation

In order to estimate the test error rate of a learning method we can randomly split the data into two parts, and use one part for training the method, and the other part for testing the obtained model. This is the **validation set approach**. The idea of **cross-validation (CV)** is to also switch the roles of the two parts of data, and use the second part for training the method, and the first part for testing or validating the model. Then we have two estimates of the test error rate, and we can use their average as a more stable estimate of the test error rate.

More generally, in the ***k*-fold cross-validation** we randomly split the data into k parts. Then we fit the method k-times, and each time we use one of the k parts as test data, and the remaining $k-1$ parts as training data. Then we get k test mean squared errors $\mathrm{MSE}_1, \ldots, \mathrm{MSE}_k$, and we can use their average to obtain the k-fold CV estimate of the test error:

$$\mathrm{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{MSE}_i.$$

4.1.1 Leave-One-Out Cross-Validation

In the **Leave-one-out cross-validation (LOOCV)** we split the data into a validation set containing only one observation $\boldsymbol{x}_i, y_i$, and a training set containing the $n-1$ remaining observations. Then we fit the learning method with the training set, and at $\boldsymbol{x}_i$ we compute the estimate $\tilde{y}_i$ for the response. The test mean squared error (MSE) of this fit is

$$\mathrm{MSE}_i = (y_i - \tilde{y}_i)^2.$$

We repeat this process n times for $i = 1, 2, \ldots, n$. Then the LOOCV estimate for the test MSE is

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i.$$

In other words, LOOCV is the n-fold CV for a data set with n data points.

When the learning method is the least squares linear regression we also have

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \Big(\frac{y_i - \hat{y}_i}{1 - h_i}\Big)^2,$$

where $\hat{y}_i$ is the prediction of the response at $\boldsymbol{x}_i$ when we fit the model with all the observations, and h_i is the ith leverage of that fit. In particular note that in this case, to compute $\mathrm{CV}_{(n)}$, we only need to fit the model once. However, for an arbitrary learning method there is no similar formula, and we have to fit the model n times to compute $\mathrm{CV}_{(n)}$.

To prove the above relation, it suffices to show that

$$y_i - \tilde{y}_i = \frac{y_i - \hat{y}_i}{1 - h_i}.$$

For simplicity we assume that $i = n$. Let $\mathbf{X}$ be the matrix of observations of predictors, and $\mathbf{y}$ the vector of observations of the response. Then we have

$$\mathbf{X} = \begin{bmatrix} \tilde{\boldsymbol{X}} \\ \boldsymbol{x}_n^\intercal \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} \mathbf{y}_{\tilde{n}} \\ y_n \end{bmatrix},$$

where $\boldsymbol{x}_n = \begin{bmatrix} 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}^\intercal$, $\tilde{\boldsymbol{X}}$ is the matrix consisting of the first $n-1$ rows of $\mathbf{X}$, and $\mathbf{y}_{\tilde{n}}$ is the vector consisting of the first $n-1$ components of $\mathbf{y}$. Let $\tilde{\mathbf{y}}, \hat{\mathbf{y}}$ be the predictions of $\mathbf{y}$ based on $n-1$ and n observations respectively. We have

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{\mathbf{y}}_{\tilde{n}} \\ \hat{y}_n \end{bmatrix} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = \begin{bmatrix} \tilde{\boldsymbol{X}} \\ \boldsymbol{x}_n^\intercal \end{bmatrix} (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = \begin{bmatrix} \tilde{\boldsymbol{X}}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} \\ \boldsymbol{x}_n^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} \end{bmatrix}.$$

We also have $\mathbf{X}^\intercal\mathbf{y} = \begin{bmatrix} \tilde{\boldsymbol{X}}^\intercal & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \mathbf{y}_{\tilde{n}} \\ y_n \end{bmatrix} = \tilde{\boldsymbol{X}}^\intercal\mathbf{y}_{\tilde{n}} + y_n\boldsymbol{x}_n$. Hence we get

$$\begin{aligned} \tilde{\mathbf{y}} = \begin{bmatrix} \tilde{\mathbf{y}}_{\tilde{n}} \\ \tilde{y}_n \end{bmatrix} &= \begin{bmatrix} \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^\intercal\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^\intercal\mathbf{y}_{\tilde{n}} \\ \boldsymbol{x}_n^\intercal(\tilde{\boldsymbol{X}}^\intercal\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^\intercal\mathbf{y}_{\tilde{n}} \end{bmatrix} \\ &= \mathbf{X}(\tilde{\boldsymbol{X}}^\intercal\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^\intercal\mathbf{y}_{\tilde{n}} = \mathbf{X}(\tilde{\boldsymbol{X}}^\intercal\tilde{\boldsymbol{X}})^{-1}(\mathbf{X}^\intercal\mathbf{y} - y_n\boldsymbol{x}_n). \qquad (*) \end{aligned}$$

Therefore we only need to compute $(\tilde{\boldsymbol{X}}^\intercal\tilde{\boldsymbol{X}})^{-1}$. However note that

$$\mathbf{X}^\intercal\mathbf{X} = \begin{bmatrix} \tilde{\boldsymbol{X}}^\intercal & \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{X}} \\ \boldsymbol{x}_n^\intercal \end{bmatrix} = \tilde{\boldsymbol{X}}^\intercal\tilde{\boldsymbol{X}} + \boldsymbol{x}_n\boldsymbol{x}_n^\intercal.$$

So $\mathbf{X}^\intercal\mathbf{X} - \tilde{\boldsymbol{X}}^\intercal\tilde{\boldsymbol{X}}$ is equal to the rank-one matrix $\boldsymbol{x}_n\boldsymbol{x}_n^\intercal$. This enables us to compute the of inverse of $\tilde{\boldsymbol{X}}^\intercal\tilde{\boldsymbol{X}}$.

More generally, suppose A, B are invertible matrices whose difference is a rank-one matrix, i.e. $A - B = uv^\intercal$, where u, v are vectors. Then we have

$$B^{-1} = (A - uv^\intercal)^{-1} = \big(A(I - A^{-1}uv^\intercal)\big)^{-1} = (I - A^{-1}uv^\intercal)^{-1}A^{-1}.$$

If we formally expand the power series of $(I - A^{-1}uv^\intercal)^{-1}$ we get

$$(I - A^{-1}uv^\intercal)^{-1} = I + A^{-1}uv^\intercal + (A^{-1}uv^\intercal)^2 + \cdots$$

But note that $(A^{-1}uv^\intercal)^2 = A^{-1}uv^\intercal A^{-1}uv^\intercal = h(A^{-1}uv^\intercal)$, where the scalar $h = v^\intercal A^{-1}u$. Thus by induction we obtain $(A^{-1}uv^\intercal)^m = h^{m-1}(A^{-1}uv^\intercal)$. Hence the formal power series becomes

$$(I - A^{-1}uv^\intercal)^{-1} = I + A^{-1}uv^\intercal(1 + h + h^2 + \cdots) = I + \frac{1}{1-h}A^{-1}uv^\intercal,$$

provided that $|h| < 1$. It is easy to see that the above matrix is in fact the inverse of $I - A^{-1}uv^\intercal$:

$$\begin{aligned}(I - A^{-1}uv^\intercal)\big(I + \frac{1}{1-h}A^{-1}uv^\intercal\big) &= I - A^{-1}uv^\intercal \\ &\quad + \frac{1}{1-h}A^{-1}uv^\intercal - \frac{1}{1-h}(A^{-1}uv^\intercal)^2 \\ &= I + A^{-1}uv^\intercal\big(-1 + \frac{1}{1-h} - \frac{h}{1-h}\big) = I.\end{aligned}$$

Note that the above computation actually shows that if A is invertible and $h \neq 1$, then B is also invertible, and we have

$$B^{-1} = (I - A^{-1}uv^\intercal)^{-1}A^{-1} = A^{-1} + \frac{1}{1-h}A^{-1}uv^\intercal A^{-1}.$$

Also notice that B^{-1} is obtained from A^{-1} by adding a rank-one matrix to it.

Now in our original problem, we have $h = \boldsymbol{x}_n^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\boldsymbol{x}_n$. However note that the projection matrix of the least square fit with all the observations is

$$\begin{aligned}\mathbf{H} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal &= \begin{bmatrix}\tilde{\boldsymbol{X}} \\ \boldsymbol{x}_n^\intercal\end{bmatrix}(\mathbf{X}^\intercal\mathbf{X})^{-1}\begin{bmatrix}\tilde{\boldsymbol{X}}^\intercal & \boldsymbol{x}_n\end{bmatrix} \\ &= \begin{bmatrix}\tilde{\boldsymbol{X}}(\mathbf{X}^\intercal\mathbf{X})^{-1}\tilde{\boldsymbol{X}}^\intercal & \tilde{\boldsymbol{X}}(\mathbf{X}^\intercal\mathbf{X})^{-1}\boldsymbol{x}_n \\ \boldsymbol{x}_n^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\tilde{\boldsymbol{X}}^\intercal & \boldsymbol{x}_n^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\boldsymbol{x}_n\end{bmatrix}.\end{aligned}$$

Therefore the leverage of the nth observation in this fit is

$$h_n = \mathbf{H}_{n,n} = \boldsymbol{x}_n^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\boldsymbol{x}_n = h.$$

Hence if the leverage $h_n \neq 1$ then $\tilde{\boldsymbol{X}}^\intercal \tilde{\boldsymbol{X}}$ is invertible too, and we have

$$(\tilde{\boldsymbol{X}}^\intercal \tilde{\boldsymbol{X}})^{-1} = (\mathbf{X}^\intercal \mathbf{X})^{-1} + \frac{1}{1-h_n}(\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n \boldsymbol{x}_n^\intercal (\mathbf{X}^\intercal \mathbf{X})^{-1}.$$

If we use this formula in the equation $(*)$, we get

$$\begin{aligned}
\tilde{\mathbf{y}} &= \mathbf{X}(\tilde{\boldsymbol{X}}^\intercal \tilde{\boldsymbol{X}})^{-1}(\mathbf{X}^\intercal \mathbf{y} - y_n \boldsymbol{x}_n) \\
&= \mathbf{X}\Big[(\mathbf{X}^\intercal \mathbf{X})^{-1} + \frac{1}{1-h_n}(\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n \boldsymbol{x}_n^\intercal (\mathbf{X}^\intercal \mathbf{X})^{-1}\Big](\mathbf{X}^\intercal \mathbf{y} - y_n \boldsymbol{x}_n) \\
&= \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y} + \frac{1}{1-h_n}\big(\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n\big)\big(\boldsymbol{x}_n^\intercal (\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y}\big) \\
&\quad - y_n \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n - \frac{y_n}{1-h_n}\big(\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n\big)\big(\boldsymbol{x}_n^\intercal (\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n\big) \\
&= \hat{\mathbf{y}} + \big(\frac{\hat{y}_n}{1-h_n} - y_n - \frac{y_n h_n}{1-h_n}\big)\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n = \hat{\mathbf{y}} + \big(\frac{\hat{y}_n - y_n}{1-h_n}\big)\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n.
\end{aligned}$$

If we subtract both sides of the above equality from $\mathbf{y}$, and look at the nth component of the resulting equation, we obtain

$$\begin{aligned}
y_n - \tilde{y}_n &= y_n - \hat{y}_n - \big(\frac{\hat{y}_n - y_n}{1-h_n}\big)\boldsymbol{x}_n^\intercal (\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n \\
&= y_n - \hat{y}_n - \big(\frac{\hat{y}_n - y_n}{1-h_n}\big)h_n = y_n - \hat{y}_n\big(1 - \frac{-h_n}{1-h_n}\big) = \frac{y_n - \hat{y}_n}{1-h_n},
\end{aligned}$$

as desired. As a consequence we get

$$|y_n - \hat{y}_n| \leq \frac{n-1}{n}|y_n - \tilde{y}_n| \leq |y_n - \tilde{y}_n|,$$

since $0 \leq 1 - h_n \leq 1 - \frac{1}{n}$. In other words, adding an observation at some point improves the linear least squares fit prediction at that point.

4.1.2 Cook's Distance

Let us further explore the above relation between $\tilde{\mathbf{y}}, \hat{\mathbf{y}}$. We have already shown that

$$\hat{\mathbf{y}} - \tilde{\mathbf{y}} = \Big(\frac{y_n - \hat{y}_n}{1-h_n}\Big)\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n.$$

Hence

$$\begin{aligned}
\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2 &= (\hat{\mathbf{y}} - \tilde{\mathbf{y}})^\intercal (\hat{\mathbf{y}} - \tilde{\mathbf{y}}) \\
&= \Big(\frac{y_n - \hat{y}_n}{1-h_n}\Big)^2 \big(\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n\big)^\intercal \big(\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\boldsymbol{x}_n\big)
\end{aligned}$$

$$= \left(\frac{y_n - \hat{y}_n}{1 - h_n}\right)^2 \boldsymbol{x}_n^\intercal (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{X} (\mathbf{X}^\intercal \mathbf{X})^{-1} \boldsymbol{x}_n$$
$$= \left(\frac{y_n - \hat{y}_n}{1 - h_n}\right)^2 \boldsymbol{x}_n^\intercal (\mathbf{X}^\intercal \mathbf{X})^{-1} \boldsymbol{x}_n = \frac{(y_n - \hat{y}_n)^2 h_n}{(1 - h_n)^2} = \frac{h_n}{1 - h_n} \sigma_\epsilon^2 t_n^2,$$

where $t_n = \frac{e_n}{\sigma_\epsilon \sqrt{1-h_n}} = \frac{y_n - \hat{y}_n}{\sigma_\epsilon \sqrt{1-h_n}}$ is the studentized residual. The quantity

$$D_n = \frac{\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2}{(p+1)\sigma_\epsilon^2} = \frac{e_n^2}{(p+1)\sigma_\epsilon^2} \frac{h_n}{(1-h_n)^2} = \frac{t_n^2}{p+1} \frac{h_n}{1-h_n}$$

is called the **Cook's distance** of the nth observation (p is the number of predictors). It measures the effect of removal of the nth observation on the predictions of the model. We can similarly compute the Cook's distance D_i of the other observations. Note that the Cook's distance is an increasing function of both the studentized residual and the leverage.

4.2 The Bootstrap

Suppose the probability distribution $\mathbb{P}$ describes a population. Also suppose that

$$(x_1, y_1), \ldots, (x_n, y_n)$$

is a sample from this population. Then we have the following estimate of $\mathbb{P}$:

$$\hat{\mathbb{P}}(x, y) = \begin{cases} \frac{1}{n} & (x, y) = (x_1, y_1), \\ \frac{1}{n} & (x, y) = (x_2, y_2), \\ \vdots & \vdots \\ \frac{1}{n} & (x, y) = (x_n, y_n), \\ 0 & \text{otherwise.} \end{cases}$$

Therefore we can use $\hat{\mathbb{P}}$ to estimate any quantity that we would be able to compute using $\mathbb{P}$. In particular, we can use $\hat{\mathbb{P}}$ to obtain new samples from the population. These new samples can for example be used to estimate test error rates, or the standard error of a coefficient in a parametric learning method.

Now note that when we form a new sample, the $\hat{\mathbb{P}}$-probability of a new data point (x, y) being in this new sample is zero; so our new sample will only contain some of the points (x_i, y_i), and under $\hat{\mathbb{P}}$ these points are all equally likely to appear in the new sample. In addition, note that when we form this new sample using $\hat{\mathbb{P}}$, in each step we randomly choose one of the points, say (x_j, y_j), with probability $\frac{1}{n}$. But in the next step we also do the same thing, and we might again randomly choose the same data point (x_j, y_j) as in the previous step. Therefore forming a

new sample using the probability distribution $\hat{\mathbb{P}}$ is the same as sampling from the data set

$$(x_1, y_1), \ldots, (x_n, y_n)$$

with **replacement**. These new samples are called **bootstrap** samples.

4.3 C_p Statistic and Adjusted R^2

Suppose we have used the observations $\mathbf{X}, \mathbf{y}$ to train a linear regression model. Also suppose that the relation between $\mathbf{X}, \mathbf{y}$ in the population is linear, i.e.

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

where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\intercal$ is a vector of i.i.d random variables. Remember that the coefficients of the linear model satisfy

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon}.$$

Now let us make another set of observations in the population at the same predictor values. These new observations will have different response values due to the random noise ϵ. We denote these new response values by $\mathbf{y}'$. Then we have

$$\mathbf{y}' = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}',$$

where $\boldsymbol{\epsilon}' = [\epsilon_1', \ldots, \epsilon_n']^\intercal$ is independent from $\boldsymbol{\epsilon}$. Since the predicted value is determined by the value of predictors, the prediction of the linear model at both sets of observations $\mathbf{X}, \mathbf{y}$ and $\mathbf{X}, \mathbf{y}'$ is the same, and is given by $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

Let us compute the expected prediction error of the model at $\mathbf{X}, \mathbf{y}'$. Recall that for a random variable like Z we have $\mathrm{Var}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2$. Hence we have

$$\begin{aligned}\mathbb{E}[(y_i' - \hat{y}_i)^2] &= \mathrm{Var}[y_i' - \hat{y}_i] + \big(\mathbb{E}[y_i' - \hat{y}_i]\big)^2 \\ &= \mathrm{Var}[y_i'] + \mathrm{Var}[\hat{y}_i] - 2\mathrm{Cov}[y_i', \hat{y}_i] + \big(\mathbb{E}[y_i'] - \mathbb{E}[\hat{y}_i]\big)^2.\end{aligned}$$

However note that $y_i' = \boldsymbol{x}_i^\intercal\boldsymbol{\beta} + \epsilon_i'$ and $y_i = \boldsymbol{x}_i^\intercal\boldsymbol{\beta} + \epsilon_i$ (where $\boldsymbol{x}_i^\intercal$ is the ith row of $\mathbf{X}$) are identically distributed and independent. Because ϵ_i, ϵ_i' are independent and have the same distribution as ϵ, and $\boldsymbol{x}_i, \boldsymbol{\beta}$ are constants. In addition, for similar reasons y_i' and

$$\hat{y}_i = \boldsymbol{x}_i^\intercal\hat{\boldsymbol{\beta}} = \boldsymbol{x}_i^\intercal\boldsymbol{\beta} + \boldsymbol{x}_i^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon}$$

are also independent. Thus $\mathrm{Cov}[y_i', \hat{y}_i] = 0$. Therefore we get

$$\begin{aligned}
\mathbb{E}[(y_i' - \hat{y}_i)^2] &= \mathrm{Var}[y_i'] + \mathrm{Var}[\hat{y}_i] + \big(\mathbb{E}[y_i'] - \mathbb{E}[\hat{y}_i]\big)^2 \\
&= \mathrm{Var}[y_i] + \mathrm{Var}[\hat{y}_i] + \big(\mathbb{E}[y_i] - \mathbb{E}[\hat{y}_i]\big)^2 \\
&= \mathrm{Var}[y_i] + \mathrm{Var}[\hat{y}_i] - 2\mathrm{Cov}[y_i, \hat{y}_i] + 2\mathrm{Cov}[y_i, \hat{y}_i] + \big(\mathbb{E}[y_i - \hat{y}_i]\big)^2 \\
&= \mathrm{Var}[y_i - \hat{y}_i] + \big(\mathbb{E}[y_i - \hat{y}_i]\big)^2 + 2\mathrm{Cov}[y_i, \hat{y}_i] \\
&= \mathbb{E}[(y_i - \hat{y}_i)^2] + 2\mathrm{Cov}[y_i, \hat{y}_i].
\end{aligned}$$

Hence the expected prediction MSE at $\mathbf{X}, \mathbf{y}'$ is

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

Note that $\frac{1}{n}\sum_{i\le n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\mathrm{RSS}$ is the training MSE of the model.

On the other hand we have

$$\mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] = \mathbf{X}\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{X}\boldsymbol{\beta}, \qquad \mathbb{E}[\hat{\mathbf{y}}] = \mathbb{E}[\mathbf{X}\hat{\boldsymbol{\beta}}] = \mathbf{X}\,\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbf{X}\boldsymbol{\beta}.$$

Therefore the covariance matrix of $\mathbf{y}, \hat{\mathbf{y}}$ is

$$\begin{aligned}
\mathbb{E}\big[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\hat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^\intercal\big] &= \mathbb{E}\big[\boldsymbol{\epsilon}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})^\intercal\big] \\
&= \mathbb{E}\big[\boldsymbol{\epsilon}(\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^\intercal\big] = \mathbb{E}\big[\boldsymbol{\epsilon}\big(\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{\epsilon}\big)^\intercal\big] \\
&= \mathbb{E}\big[\boldsymbol{\epsilon}(\mathbf{H}\boldsymbol{\epsilon})^\intercal\big] = \mathbb{E}\big[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal\mathbf{H}\big] = \mathbb{E}\big[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\intercal\big]\mathbf{H} = \sigma_\epsilon^2 I\,\mathbf{H} = \sigma_\epsilon^2\,\mathbf{H},
\end{aligned}$$

where $\mathbf{H}$ is the projection matrix. So the expected prediction MSE at $\mathbf{X}, \mathbf{y}'$ is

$$\begin{aligned}
\mathbb{E}\Big[\frac{1}{n}\sum_{i\le n}(y_i' - \hat{y}_i)^2\Big] &= \mathbb{E}\Big[\frac{1}{n}\sum_{i\le n}(y_i - \hat{y}_i)^2\Big] + \frac{2}{n}\sum_{i\le n}\mathrm{Cov}[y_i, \hat{y}_i] \\
&= \mathbb{E}\Big[\frac{1}{n}\mathrm{RSS}\Big] + \frac{2}{n}\sum_{i\le n}\sigma_\epsilon^2\,\mathbf{H}_{ii} \\
&= \mathbb{E}\Big[\frac{1}{n}\mathrm{RSS}\Big] + \frac{2}{n}\sigma_\epsilon^2\,\mathrm{tr}(\mathbf{H}) = \mathbb{E}\Big[\frac{1}{n}\mathrm{RSS}\Big] + \frac{2}{n}\sigma_\epsilon^2 d,
\end{aligned}$$

where d is the number of coefficients in the linear model (i.e. $d = p + 1$, where p is the number of predictors). Now let $\hat{\sigma}^2$ be an unbiased estimator of σ_ϵ^2 (for example we can use RSE^2). Then we have

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i\le n}(y_i' - \hat{y}_i)^2\Big] = \mathbb{E}\Big[\frac{1}{n}\mathrm{RSS}\Big] + \frac{2}{n}\sigma_\epsilon^2 d = \mathbb{E}\Big[\frac{1}{n}\mathrm{RSS} + \frac{2d}{n}\hat{\sigma}^2\Big].$$

Therefore

$$C_p = \frac{1}{n}\big(\mathrm{RSS} + 2d\hat{\sigma}^2\big)$$

is an unbiased estimator of the expected prediction MSE at $\mathbf{X}, \mathbf{y}'$. Note that if we add more predictors to the model, RSS will decrease, but $2d\hat{\sigma}^2$ increases. Hence unlike the training MSE, C_p does not necessarily decrease as we add more predictors to the model. This property of C_p, and the fact that it is an estimator of the prediction MSE, makes it a suitable statistic for comparing models with different numbers of predictors.

Another quantity for comparing models with different numbers of predictors is the **adjusted R^2**, defined by

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}.$$

Equivalently we have

$$1 - R^2_{\text{adj}} = (1 - R^2)\frac{n-1}{n-p-1}.$$

The factor $\frac{n-1}{n-p-1}$ is included in order to adjust the effect of having more predictors on the R^2 statistic. When we add a predictor to the model, we know that RSS decreases. However, $n-p-1$ decreases too. So R^2_{adj} does not necessarily increase, in contrast to R^2.

Now note that when we change the number of predictors, $\text{TSS}/(n-1)$ does not change. Hence the change of R^2_{adj} is only due to the change in $\text{RSS}/(n-p-1)$. Let us denote the number of coefficients by d, i.e. $d = p+1$. Then for some unknown λ we have

$$\frac{\text{RSS}}{n-d} = \frac{1}{n}\big(\text{RSS} + \lambda\big).$$

Let us compute λ. We have

$$\lambda = \frac{n\,\text{RSS}}{n-d} - \text{RSS} = \text{RSS}\big(\frac{n}{n-d} - 1\big) = \text{RSS}\frac{d}{n-d} = d\hat{\sigma}^2,$$

where $\hat{\sigma}^2 = \text{RSS}/(n-d)$ is the RSE, which we know is an unbiased estimator of σ^2_ϵ. Therefore we get

$$\begin{aligned} R^2_{\text{adj}} &= 1 - \frac{\text{RSS}/(n-d)}{\text{TSS}/(n-1)} \\ &= 1 - \frac{(\text{RSS} + d\hat{\sigma}^2)/n}{\text{TSS}/(n-1)} = 1 - \frac{C_p - \frac{d}{n}\hat{\sigma}^2}{\text{TSS}/(n-1)}. \end{aligned}$$

But C_p is an unbiased estimator of the expected prediction MSE. Thus $C_p - \frac{d}{n}\hat{\sigma}^2$ tends to underestimate the expected prediction MSE. Hence R^2_{adj} tends to overestimate the prediction accuracy of the model. However, as we have said, it can be used to compare models with different numbers of predictors.

4.4 Regularization

Suppose we have a set of observed data $\mathbf{X}, \mathbf{y}$, and we want to use this data to find the estimate $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ for $\mathbf{y}$. In linear regression we find $\hat{\boldsymbol{\beta}}$ by minimizing the RSS. The idea of **regularization** is to constrain $\hat{\boldsymbol{\beta}}$ so that it cannot take large values, and as a result its variance decreases. More precisely, instead of minimizing the RSS we minimize

$$\mathrm{RSS}(\hat{\boldsymbol{\beta}}) + \lambda J(\hat{\boldsymbol{\beta}}) = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \lambda J(\hat{\boldsymbol{\beta}}),$$

where $J : \mathbb{R}^{p+1} \to \mathbb{R}$ is a given function, and λ is a given nonnegative number. In the process of minimization, the penalty term $\lambda J(\hat{\boldsymbol{\beta}})$ forces $\hat{\boldsymbol{\beta}}$ to stay closer to the origin compared to the case where there is no penalty.

First let us show that minimizing $\mathrm{RSS}+\lambda J$ over $\mathbb{R}^{p+1}$ is equivalent to minimizing RSS over the set $\{J \le s\}$ for some $s \ge 0$. Let $\hat{\boldsymbol{\beta}}$ be the minimizer of $\mathrm{RSS} + \lambda J$ over $\mathbb{R}^{p+1}$. Set $s := J(\hat{\boldsymbol{\beta}})$. Let $\tilde{\boldsymbol{\beta}}$ be the minimizer of RSS over the set $\{J \le s\}$. Note that $\hat{\boldsymbol{\beta}} \in \{J \le s\}$. Hence we must have

$$\mathrm{RSS}(\tilde{\boldsymbol{\beta}}) \le \mathrm{RSS}(\hat{\boldsymbol{\beta}}).$$

On the other hand we know that $\mathrm{RSS}(\hat{\boldsymbol{\beta}}) + \lambda J(\hat{\boldsymbol{\beta}}) \le \mathrm{RSS}(\tilde{\boldsymbol{\beta}}) + \lambda J(\tilde{\boldsymbol{\beta}})$. Combining these two inequalities we get

$$\begin{aligned}\mathrm{RSS}(\hat{\boldsymbol{\beta}}) + \lambda J(\hat{\boldsymbol{\beta}}) \le \mathrm{RSS}(\tilde{\boldsymbol{\beta}}) + \lambda J(\tilde{\boldsymbol{\beta}}) &\le \mathrm{RSS}(\hat{\boldsymbol{\beta}}) + \lambda J(\tilde{\boldsymbol{\beta}}) \\ &\le \mathrm{RSS}(\hat{\boldsymbol{\beta}}) + \lambda s = \mathrm{RSS}(\hat{\boldsymbol{\beta}}) + \lambda J(\hat{\boldsymbol{\beta}}).\end{aligned}$$

Therefore we have $\mathrm{RSS}(\hat{\boldsymbol{\beta}}) + \lambda J(\hat{\boldsymbol{\beta}}) = \mathrm{RSS}(\tilde{\boldsymbol{\beta}}) + \lambda J(\tilde{\boldsymbol{\beta}})$. Thus if we further assume that $\mathrm{RSS}+\lambda J$ has a unique minimizer (which for example is the case when $\mathrm{RSS}+\lambda J$ is strictly convex), we obtain $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$, as desired.

4.4.1 Ridge Regression and Lasso

If we assume that J is differentiable, we can try to find the minimizer $\hat{\boldsymbol{\beta}}$ by differentiation. Let $\boldsymbol{\gamma} \in \mathbb{R}^{p+1}$ be an arbitrary vector. Then the directional derivative of $\mathrm{RSS} + \lambda J$ at $\hat{\boldsymbol{\beta}}$ in the direction of $\boldsymbol{\gamma}$ must be zero. Hence we must have

$$\begin{aligned}0 &= \frac{d}{dt}\Big|_{t=0} \big(\mathrm{RSS}(\hat{\boldsymbol{\beta}} + t\boldsymbol{\gamma}) + \lambda J(\hat{\boldsymbol{\beta}} + t\boldsymbol{\gamma})\big) \\ &= \frac{d}{dt}\Big|_{t=0} \big((\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - t\mathbf{X}\boldsymbol{\gamma})^{\intercal}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - t\mathbf{X}\boldsymbol{\gamma}) + \lambda J(\hat{\boldsymbol{\beta}} + t\boldsymbol{\gamma})\big)\end{aligned}$$

$$
\begin{aligned}
&= \big(\frac{d}{dt}\Big|_{t=0}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}-t\mathbf{X}\boldsymbol{\gamma})^{\intercal}\big)(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\
&\qquad + (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^{\intercal}\big(\frac{d}{dt}\Big|_{t=0}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}-t\mathbf{X}\boldsymbol{\gamma})\big) + \lambda\frac{d}{dt}\Big|_{t=0}J(\hat{\boldsymbol{\beta}}+t\boldsymbol{\gamma}) \\
&= (-\mathbf{X}\boldsymbol{\gamma})^{\intercal}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) + (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^{\intercal}(-\mathbf{X}\boldsymbol{\gamma}) + \lambda D_{\boldsymbol{\gamma}}J(\hat{\boldsymbol{\beta}}) \\
&= -\boldsymbol{\gamma}^{\intercal}\mathbf{X}^{\intercal}\mathbf{y} + \boldsymbol{\gamma}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y}^{\intercal}\mathbf{X}\boldsymbol{\gamma} + \hat{\boldsymbol{\beta}}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\boldsymbol{\gamma} + \lambda\boldsymbol{\gamma}^{\intercal}DJ(\hat{\boldsymbol{\beta}}) \\
&= -2\boldsymbol{\gamma}^{\intercal}\mathbf{X}^{\intercal}\mathbf{y} + 2\boldsymbol{\gamma}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\hat{\boldsymbol{\beta}} + \lambda\boldsymbol{\gamma}^{\intercal}DJ(\hat{\boldsymbol{\beta}}) \qquad (\text{since } a^{\intercal}=a \text{ for a } 1\times 1 \text{ matrix}) \\
&= 2\boldsymbol{\gamma}^{\intercal}\big(-\mathbf{X}^{\intercal}\mathbf{y} + \mathbf{X}^{\intercal}\mathbf{X}\hat{\boldsymbol{\beta}} + \frac{1}{2}\lambda DJ(\hat{\boldsymbol{\beta}})\big).
\end{aligned}
$$

Therefore we get

$$\mathbf{X}^{\intercal}\mathbf{X}\hat{\boldsymbol{\beta}} + \frac{1}{2}\lambda DJ(\hat{\boldsymbol{\beta}}) = \mathbf{X}^{\intercal}\mathbf{y}, \qquad (*)$$

since $\boldsymbol{\gamma}$ is arbitrary.

Now suppose $J(\hat{\boldsymbol{\beta}}) = \hat{\beta}_1^2+\cdots+\hat{\beta}_p^2 = \hat{\boldsymbol{\beta}}^{\intercal}\tilde{I}\hat{\boldsymbol{\beta}}$, where $\tilde{I}$ is the diagonal matrix whose diagonal entries are all 1 except the first diagonal entry which is 0. This method of regularization is known as **ridge regression**, and its corresponding coefficient is denoted by $\hat{\boldsymbol{\beta}}_\lambda^R$. In this case the equation $(*)$ becomes

$$\mathbf{X}^{\intercal}\mathbf{y} = \mathbf{X}^{\intercal}\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda^R + \lambda\tilde{I}\hat{\boldsymbol{\beta}}_\lambda^R = (\mathbf{X}^{\intercal}\mathbf{X}+\lambda\tilde{I})\hat{\boldsymbol{\beta}}_\lambda^R.$$

Note that $\det(\mathbf{X}^{\intercal}\mathbf{X}+\lambda\tilde{I})$ is a polynomial in λ; so it has finitely many roots. Thus the matrix $\mathbf{X}^{\intercal}\mathbf{X}+\lambda\tilde{I}$ is invertible for almost all values of λ (even when $p>n$), and in this case we have

$$\hat{\boldsymbol{\beta}}_\lambda^R = (\mathbf{X}^{\intercal}\mathbf{X}+\lambda\tilde{I})^{-1}\mathbf{X}^{\intercal}\mathbf{y}.$$

As a result note that we can find $\hat{\boldsymbol{\beta}}_\lambda^R$ for all such values of λ simultaneously, since the entries of $(\mathbf{X}^{\intercal}\mathbf{X}+\lambda\tilde{I})^{-1}$ are rational functions of λ, and can be calculated by standard algorithms for solving linear systems.

Next suppose $J(\hat{\boldsymbol{\beta}}) = |\hat{\beta}_1|+\cdots+|\hat{\beta}_p|$. This method of regularization is known as the **lasso**. Note that in this case J is not always differentiable; we have

$$D_jJ = \begin{cases} \operatorname{sign}(\hat{\beta}_j) & \hat{\beta}_j\neq 0, \\ \text{does not exist} & \hat{\beta}_j = 0, \end{cases} = \begin{cases} 1 & \hat{\beta}_j > 0, \\ -1 & \hat{\beta}_j < 0, \\ \text{does not exist} & \hat{\beta}_j = 0. \end{cases}$$

Hence when we look for the minimizer of $\mathrm{RSS}+\lambda J$, in addition to the points where the derivative vanishes, we also have to check the points of nondifferentiability, i.e. the points which have some zero components. The possibility that these points with some zero components can be the minimizer implies that the lasso can perform **feature selection** too.

4.4.2 Regularization from Bayesian viewpoint

Suppose the probability measure $\mathbb{P}$ describes the distribution of some quantity X among a population. We assume that $\mathbb{P}$ belongs to a parametric family of probability measures $\mathbb{P}_\theta$, but we do not know which value of θ gives us $\mathbb{P}$. In classical statistics X is random, but θ is a fixed unknown parameter (which we try to estimate). In contrast, in **Bayesian statistics** both X, θ are considered to be random variables. Let $f(x|\theta)$ denote the probability density (mass) function of $\mathbb{P}_\theta$, which is the distribution of X given θ. As before, $f(x|\theta)$ is called the **likelihood function**. Let $\pi(\theta)$ be the **prior density** of θ, i.e. the density of θ before we know anything about X. Then the joint probability density of X, θ is $\pi(\theta)f(x|\theta)$ (note that unlike X, we use θ to denote both a random variable and its values). Hence the marginal density of X (before knowing θ) is

$$f(x) := \int_\Theta \pi(\vartheta) f(x|\vartheta)\, d\vartheta,$$

where Θ is the set of all possible values of θ. Thus by Bayes' theorem, the **posterior density** of θ, given the observation $X = x$, is

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{f(x)}.$$

This relation is usually written as

$$\pi(\theta|x) \propto \pi(\theta)f(x|\theta),$$

where the constant of proportionality, i.e. $f(x)$, can depend on x but not on θ. In other words we have

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

The **posterior mode** $\hat{\theta}$ is the value of θ that maximizes the posterior $\pi(\theta|x)$, i.e. $\hat{\theta}$ is the most likely value of θ, given the observed data x. We have

$$\hat{\theta} = \arg\max_\theta \pi(\theta|x) = \arg\max_\theta \log \pi(\theta|x),$$

since log is an increasing function. Hence

$$\begin{aligned}
\hat{\theta} = \arg\max_\theta \log \pi(\theta|x) &= \arg\max_\theta \log\Big(\frac{\pi(\theta)f(x|\theta)}{f(x)}\Big) \\
&= \arg\max_\theta \big(\log \pi(\theta) + \log f(x|\theta) - \log f(x)\big) \\
&= \arg\max_\theta \big(\log \pi(\theta) + \log f(x|\theta)\big).
\end{aligned}$$

Where the last equality holds since $f(x)$ does not depend on θ. In comparison, recall that $\arg\max_\theta \log f(x|\theta)$ is the maximum likelihood estimator of θ.

Now let us consider the regularization in linear regression. Suppose

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon,$$

where ϵ is a normally distributed noise with zero mean and variance σ^2. (For simplicity we assume that $\beta_0 = 0$, and all variables have zero mean.) Suppose we have n observations of $Y, X_1, \ldots, X_p$, which satisfy

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i.$$

And ϵ_is are i.i.d. random variables with the same distribution as ϵ. In this case θ is the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ (and y_i plays the role of x in the above discussion, while we treat $x_{i1}, \ldots, x_{ip}$ as constants). If x_{ij}s and β_js are fixed, then $\sum \beta_j x_{ij}$ is constant; so y_i is normally distributed with mean $\sum \beta_j x_{ij}$, and variance σ^2, i.e. we have

$$f(y_i|\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_i - \sum \beta_j x_{ij})^2}{\sigma^2}\Big).$$

(Be careful not to confuse the mathematical constant π with the prior or posterior densities!) Then since y_is are independent we have

$$f(\mathbf{y}|\boldsymbol{\beta}) = f(y_1, \ldots, y_n|\boldsymbol{\beta}) = \prod_{i\le n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(y_i - \sum \beta_j x_{ij})^2}{\sigma^2}\Big).$$

Hence

$$\log f(\mathbf{y}|\boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{\sigma^2}\sum_{i\le n}\big(y_i - \sum_{j\ge 1}\beta_j x_{ij}\big)^2.$$

Now let us assume that $\pi(\boldsymbol{\beta}) = \prod_{j\ge 1} \frac{1}{\sqrt{2\pi\tau^2}} \exp\big(-\beta_j^2/\tau^2\big)$, i.e. we assume that $\pi(\beta_j)$ is normal with zero mean and variance τ^2, and β_js are independent. Then we get

$$\log \pi(\boldsymbol{\beta}) = -\frac{p}{2}\log(2\pi\tau^2) - \frac{1}{\tau^2}\sum_{j\ge 1}\beta_j^2.$$

Therefore we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\max_{\boldsymbol{\beta}} \big(\log \pi(\boldsymbol{\beta}) + \log f(\mathbf{y}|\boldsymbol{\beta})\big) \\
&= \arg\max_{\boldsymbol{\beta}} \Big(- \frac{1}{\sigma^2}\sum \big(y_i - \sum \beta_j x_{ij}\big)^2 - \frac{1}{\tau^2}\sum \beta_j^2\Big) \\
&= \arg\min_{\boldsymbol{\beta}} \Big(\sum \big(y_i - \sum \beta_j x_{ij}\big)^2 + \frac{\sigma^2}{\tau^2}\sum \beta_j^2\Big) \\
&= \arg\min_{\boldsymbol{\beta}} \big(\mathrm{RSS}(\boldsymbol{\beta}) + \lambda \sum \beta_j^2\big).
\end{aligned}$$

Note that in the second and third equalities above we are using the fact that σ, τ are (positive) constants, and therefore canceling them or factoring them out does not alter the extremum point. Thus we have shown that when the prior density of β_js is normal with zero mean and variance τ^2, the posterior mode $\hat{\boldsymbol{\beta}}$ is also the ridge regression coefficient corresponding to $\lambda = \sigma^2/\tau^2$.

Next let us assume that $\pi(\boldsymbol{\beta}) = \prod_{j\geq 1} \frac{1}{2b} \exp\big(-|\beta_j|/b\big)$, i.e. we assume that $\pi(\beta_j)$ is Laplace distribution with zero mean and scale parameter b, and β_js are independent. Then we get

$$\log \pi(\boldsymbol{\beta}) = -p \log(2b) - \frac{1}{b} \sum_{j\geq 1} |\beta_j|.$$

Therefore we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\max_{\boldsymbol{\beta}} \big(\log \pi(\boldsymbol{\beta}) + \log f(\mathbf{y}|\boldsymbol{\beta})\big) \\
&= \arg\max_{\boldsymbol{\beta}} \Big(-\frac{1}{\sigma^2} \sum \big(y_i - \sum \beta_j x_{ij}\big)^2 - \frac{1}{b} \sum |\beta_j|\Big) \\
&= \arg\min_{\boldsymbol{\beta}} \Big(\sum \big(y_i - \sum \beta_j x_{ij}\big)^2 + \frac{\sigma^2}{b} \sum |\beta_j|\Big) \\
&= \arg\min_{\boldsymbol{\beta}} \big(\mathrm{RSS}(\boldsymbol{\beta}) + \lambda \sum |\beta_j|\big).
\end{aligned}$$

Thus we have shown that when the prior density of β_js is Laplace distribution with zero mean and scale parameter b, the posterior mode $\hat{\boldsymbol{\beta}}$ is also the lasso coefficient corresponding to $\lambda = \sigma^2/b$.

4.5 Principal Components Analysis

Let $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_p \end{bmatrix}$ be the matrix of n observations of $X_1, \dots, X_p$ (note that we do not include the column of 1s). We assume that the observations are shifted to have zero mean, i.e. $\bar{x}_j = 0$ for each j. Suppose we want to find a linear combination of $\mathbf{x}_1, \dots, \mathbf{x}_p$ which has the largest variance, i.e. we want to determine the coefficients $\boldsymbol{\phi}_1 = \begin{bmatrix} \phi_{11} & \dots & \phi_{p1} \end{bmatrix}^\intercal$ so that

$$\mathbf{z}_1 = \phi_{11}\mathbf{x}_1 + \cdots + \phi_{p1}\mathbf{x}_p = \mathbf{X}\boldsymbol{\phi}_1$$

has the maximum variance among all linear combinations of $\mathbf{x}_1, \dots, \mathbf{x}_p$. In other words, we want to find the direction $\boldsymbol{\phi}_1$ in the feature space along which the data has the largest variance. Informally, the direction with the greatest variance contains a large amount of the information in the data, because from a statistical point of view there is little information in a variable with small variance, since that variable is

close to being a constant. Now note that if ϕ_{j1}s are allowed to be arbitrarily large, then the variance of $\mathbf{z}_1$ can be made arbitrarily large as well. Thus we require that $\phi_{11}^2 + \cdots + \phi_{p1}^2 = \|\boldsymbol{\phi}_1\|^2 = 1$, i.e. we only care about the direction determined by $\boldsymbol{\phi}_1$, and not its magnitude.

If we write the above relation in coordinates we have

$$z_{i1} = \sum_{j=1}^{p} \phi_{j1} x_{ij}.$$

Hence $\bar{z}_1 = \sum_{j=1}^{p} \phi_{j1} \bar{x}_j = 0$. Therefore the variance of z_{i1}s can be computed as follows

$$\begin{aligned}(n-1)s_{z_1}^2 = \sum_{i \le n} (z_{i1} - \bar{z}_1)^2 = \sum_{i \le n} z_{i1}^2 = \|\mathbf{z}_1\|^2 \\ = \mathbf{z}_1^\intercal \mathbf{z}_1 = (\mathbf{X}\boldsymbol{\phi}_1)^\intercal \mathbf{X}\boldsymbol{\phi}_1 = \boldsymbol{\phi}_1^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi}_1.\end{aligned}$$

But from linear algebra we know that the maximum of $\boldsymbol{\phi}_1^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi}_1$ among the vectors $\boldsymbol{\phi}_1$ with $\|\boldsymbol{\phi}_1\| = 1$ is achieved when $\boldsymbol{\phi}_1$ is an eigenvector of the symmetric matrix $\mathbf{X}^\intercal \mathbf{X}$ corresponding to its largest eigenvalue, which we call λ_1, i.e.

$$\boldsymbol{\phi}_1 = \operatorname*{arg\,max}_{\|\boldsymbol{\phi}\|=1} \boldsymbol{\phi}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi}, \quad \lambda_1 = \max_{\|\boldsymbol{\phi}\|=1} \boldsymbol{\phi}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi} \implies \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi}_1 = \lambda_1 \boldsymbol{\phi}_1.$$

(The proof is presented below.) As a result we have

$$(n-1)s_{z_1}^2 = \|\mathbf{z}_1\|^2 = \boldsymbol{\phi}_1^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi}_1 = \lambda_1 \boldsymbol{\phi}_1^\intercal \boldsymbol{\phi}_1 = \lambda_1 \|\boldsymbol{\phi}_1\|^2 = \lambda_1,$$

i.e. the sample variance of $\mathbf{z}_1$ is $\lambda_1/(n-1)$. Note that $\mathbf{X}^\intercal \mathbf{X}$ is a positive matrix, so its eigenvalues are nonnegative. And recall that by definition, the eigenvalues of $\mathbf{X}^\intercal \mathbf{X}$ are the squares of the singular values of $\mathbf{X}$. The variable $\mathbf{z}_1$ is called the **first principal component** of $\mathbf{x}_1, \ldots, \mathbf{x}_p$, and $\boldsymbol{\phi}_1$ is called the first principal component **loading** vector. Also, $z_{11}, \ldots, z_{n1}$ are known as the **scores** of the first principal component.

The **second principal component** of $\mathbf{x}_1, \ldots, \mathbf{x}_p$ is a linear combination

$$\mathbf{z}_2 = \phi_{12}\mathbf{x}_1 + \cdots + \phi_{p2}\mathbf{x}_p = \mathbf{X}\boldsymbol{\phi}_2$$

which has the largest variance among those linear combinations of $\mathbf{x}_1, \ldots, \mathbf{x}_p$ that are uncorrelated with $\mathbf{z}_1$. Note that the sample mean of $\mathbf{z}_2$, like all other linear combinations of $\mathbf{x}_1, \ldots, \mathbf{x}_p$, is zero. Thus the covariance of $\mathbf{z}_1, \mathbf{z}_2$ is equal to their inner product, because

$$(n-1)q_{z_1 z_2} = \sum_{i \le n} (z_{i1} - \bar{z}_1)(z_{i2} - \bar{z}_2) = \sum_{i \le n} z_{i1} z_{i2} = \mathbf{z}_1^\intercal \mathbf{z}_2.$$

Hence $\mathbf{z}_2$ has the largest variance among those linear combinations of $\mathbf{x}_1, \ldots, \mathbf{x}_p$ which are orthogonal to $\mathbf{z}_1$. Thus to obtain $\boldsymbol{\phi}_2$ we need to maximize $\boldsymbol{\phi}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi}$ among all vectors $\boldsymbol{\phi}$ with length one such that $\mathbf{X}\boldsymbol{\phi}$ is orthogonal to $\mathbf{X}\boldsymbol{\phi}_1 = \mathbf{z}_1$. But in this case we have

$$0 = (\mathbf{X}\boldsymbol{\phi})^\intercal \mathbf{X}\boldsymbol{\phi}_1 = \boldsymbol{\phi}^\intercal \mathbf{X}^\intercal \mathbf{X}\boldsymbol{\phi}_1 = \lambda_1 \boldsymbol{\phi}^\intercal \boldsymbol{\phi}_1.$$

Since we assume that $\mathbf{X}^\intercal \mathbf{X}$ is invertible, λ_1 is nonzero; so we must have $\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_1 = 0$, i.e. $\boldsymbol{\phi}, \boldsymbol{\phi}_1$ must be orthogonal. Therefore $\boldsymbol{\phi}_2$ maximizes $\boldsymbol{\phi}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi}$ among all vectors $\boldsymbol{\phi}$ with length one which are orthogonal to $\boldsymbol{\phi}_1$. However, from linear algebra we know this happens when (and only when) $\boldsymbol{\phi}_2$ is an eigenvector of $\mathbf{X}^\intercal \mathbf{X}$ corresponding to its second largest eigenvalue, which we call λ_2. (The proof is presented below. Also note that λ_2 can be equal to λ_1 too.)

We can similarly define the ***k*th principal component** $\mathbf{z}_k = \mathbf{X}\boldsymbol{\phi}_k$ as the linear combination of $\mathbf{x}_1, \ldots, \mathbf{x}_p$ which has the largest variance among those linear combinations of $\mathbf{x}_1, \ldots, \mathbf{x}_p$ that are uncorrelated with $\mathbf{z}_1, \ldots, \mathbf{z}_{k-1}$. Then similarly to the above, we can easily show that this is equivalent to $\boldsymbol{\phi}_k$ being the maximizer of $\boldsymbol{\phi}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi}$ among all vectors $\boldsymbol{\phi}$ with length one which are orthogonal to $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{k-1}$. And again we know this happens when (and only when) $\boldsymbol{\phi}_k$ is an eigenvector of $\mathbf{X}^\intercal \mathbf{X}$ corresponding to its kth largest eigenvalue, which we call λ_k. The reason is that the symmetric matrix $\mathbf{X}^\intercal \mathbf{X}$ has an orthonormal basis of eigenvectors, which we denote by $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p$, and we assume they are ordered to make the sequence of eigenvalues decreasing. Thus if $\boldsymbol{\phi}$ is orthogonal to $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{k-1}$, it must be a linear combination of $\boldsymbol{\phi}_k, \ldots, \boldsymbol{\phi}_p$. Hence we have

$$\boldsymbol{\phi} = (\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_k)\boldsymbol{\phi}_k + \cdots + (\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_p)\boldsymbol{\phi}_p,$$

since $\boldsymbol{\phi}_k, \ldots, \boldsymbol{\phi}_p$ is an orthonormal set. Therefore

$$\begin{aligned}
\boldsymbol{\phi}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi} &= \boldsymbol{\phi}^\intercal \mathbf{X}^\intercal \mathbf{X}\big((\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_k)\boldsymbol{\phi}_k + \cdots + (\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_p)\boldsymbol{\phi}_p\big) \\
&= \boldsymbol{\phi}^\intercal \big((\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_k)\mathbf{X}^\intercal \mathbf{X}\boldsymbol{\phi}_k + \cdots + (\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_p)\mathbf{X}^\intercal \mathbf{X}\boldsymbol{\phi}_p\big) \\
&= \boldsymbol{\phi}^\intercal \big((\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_k)\lambda_k \boldsymbol{\phi}_k + \cdots + (\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_p)\lambda_p \boldsymbol{\phi}_p\big) \\
&= \lambda_k (\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_k)^2 + \cdots + \lambda_p (\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_p)^2 \\
&\leq \lambda_k \big((\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_k)^2 + \cdots + (\boldsymbol{\phi}^\intercal \boldsymbol{\phi}_p)^2\big) = \lambda_k \|\boldsymbol{\phi}\|^2 = \lambda_k,
\end{aligned}$$

since λ_k is the largest eigenvalue among $\lambda_k, \ldots, \lambda_p$. Furthermore note that the maximum of $\boldsymbol{\phi}^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi}$ is achieved when $\boldsymbol{\phi} = \boldsymbol{\phi}_k$, because

$$\boldsymbol{\phi}_k^\intercal \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\phi}_k = \boldsymbol{\phi}_k^\intercal (\lambda_k \boldsymbol{\phi}_k) = \lambda_k \boldsymbol{\phi}_k^\intercal \boldsymbol{\phi}_k = \lambda_k \|\boldsymbol{\phi}_k\|^2 = \lambda_k.$$

Thus $\boldsymbol{\phi}_k$ is the desired maximizer.

We can continue the above process for $k \leq p$. Therefore, by construction $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p$ is an orthonormal set of eigenvectors of $\mathbf{X}^\intercal\mathbf{X}$ corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. In addition, the principal components $\mathbf{z}_1, \ldots, \mathbf{z}_p$ also form an orthogonal set, which implies they are uncorrelated, since their sample means are zero. Furthermore for every k we have

$$\begin{aligned}(n-1)s_{z_k}^2 &= \sum_{i\leq n}(z_{ik} - \bar{z}_k)^2 = \sum_{i\leq n} z_{ik}^2 = \|\mathbf{z}_k\|^2 \\ &= \mathbf{z}_k^\intercal\mathbf{z}_k = (\mathbf{X}\boldsymbol{\phi}_k)^\intercal\mathbf{X}\boldsymbol{\phi}_k = \boldsymbol{\phi}_k^\intercal\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\phi}_k = \lambda_k\boldsymbol{\phi}_k^\intercal\boldsymbol{\phi}_k = \lambda_k\|\boldsymbol{\phi}_k\|^2 = \lambda_k;\end{aligned}$$

so $s_{z_k}^2 = \lambda_k/(n-1)$.

Another interesting property is that for any $m \leq p$, $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_m$ span the closest m-dimensional linear subspace to the data. To explain this let us denote the rows of the matrix $\mathbf{X}$ by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, i.e. we have

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1^\intercal \\ \boldsymbol{x}_2^\intercal \\ \vdots \\ \boldsymbol{x}_n^\intercal \end{bmatrix}.$$

Also note that $\boldsymbol{x}_i \in \mathbb{R}^p$ is the vector of the ith observation. In addition we have

$$\mathbf{z}_k = \mathbf{X}\boldsymbol{\phi}_k = \begin{bmatrix} \boldsymbol{x}_1^\intercal \\ \vdots \\ \boldsymbol{x}_n^\intercal \end{bmatrix} \boldsymbol{\phi}_k = \begin{bmatrix} \boldsymbol{x}_1^\intercal\boldsymbol{\phi}_k \\ \vdots \\ \boldsymbol{x}_n^\intercal\boldsymbol{\phi}_k \end{bmatrix}.$$

Thus $z_{ik} = \boldsymbol{x}_i^\intercal\boldsymbol{\phi}_k$, i.e. the ith component of $\mathbf{z}_k$ is the inner product of $\boldsymbol{x}_i$ and $\boldsymbol{\phi}_k$. On the other hand, $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_m$ form an orthonormal set. Therefore the orthogonal projection of $\boldsymbol{x}_i$ on the subspace $V = \operatorname{span}(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_m)$ is given by

$$\boldsymbol{v}_i = (\boldsymbol{x}_i^\intercal\boldsymbol{\phi}_1)\boldsymbol{\phi}_1 + \cdots + (\boldsymbol{x}_i^\intercal\boldsymbol{\phi}_m)\boldsymbol{\phi}_m.$$

Furthermore note that, due to the properties of orthogonal projection, $\boldsymbol{v}_i$ and $\boldsymbol{x}_i - \boldsymbol{v}_i$ are orthogonal; so by Pythagorean theorem we have $\|\boldsymbol{x}_i\|^2 = \|\boldsymbol{v}_i\|^2 + \|\boldsymbol{x}_i - \boldsymbol{v}_i\|^2$.

Hence the sum of the squared distance of the observed data from the subspace V is

$$\begin{aligned}\sum_{i\leq n}\|\boldsymbol{x}_i - \boldsymbol{v}_i\|^2 &= \sum_{i\leq n}\|\boldsymbol{x}_i\|^2 - \sum_{i\leq n}\|\boldsymbol{v}_i\|^2 \\ &= \sum_{i\leq n}\|\boldsymbol{x}_i\|^2 - \sum_{i\leq n}\sum_{k\leq m}(\boldsymbol{x}_i^\intercal\boldsymbol{\phi}_k)^2 \qquad \text{(since } \boldsymbol{\phi}_k\text{s are orthonormal)} \\ &= \sum_{i\leq n}\|\boldsymbol{x}_i\|^2 - \sum_{k\leq m}\sum_{i\leq n}(\boldsymbol{x}_i^\intercal\boldsymbol{\phi}_k)^2 \\ &= \sum_{i\leq n}\|\boldsymbol{x}_i\|^2 - \sum_{k\leq m}\|\mathbf{z}_k\|^2 = \sum_{i\leq n}\|\boldsymbol{x}_i\|^2 - \sum_{k\leq m}\lambda_k.\end{aligned}$$

Now suppose W is another m-dimensional subspace of $\mathbb{R}^p$. Let $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_m$ be an orthonormal basis for W. Then the orthogonal projection of $\boldsymbol{x}_i$ on W is

$$\boldsymbol{w}_i = (\boldsymbol{x}_i^\intercal \boldsymbol{\psi}_1)\boldsymbol{\psi}_1 + \cdots + (\boldsymbol{x}_i^\intercal \boldsymbol{\psi}_m)\boldsymbol{\psi}_m.$$

Thus we can similarly compute the sum of the squared distance of the observed data from the subspace W:

$$\begin{aligned}
\sum_{i\le n} \|\boldsymbol{x}_i - \boldsymbol{w}_i\|^2 &= \sum_{i\le n} \|\boldsymbol{x}_i\|^2 - \sum_{i\le n} \|\boldsymbol{w}_i\|^2 \\
&= \sum_{i\le n} \|\boldsymbol{x}_i\|^2 - \sum_{i\le n}\sum_{k\le m} (\boldsymbol{x}_i^\intercal \boldsymbol{\psi}_k)^2 \\
&= \sum_{i\le n} \|\boldsymbol{x}_i\|^2 - \sum_{k\le m}\sum_{i\le n} (\boldsymbol{x}_i^\intercal \boldsymbol{\psi}_k)^2 = \sum_{i\le n} \|\boldsymbol{x}_i\|^2 - \sum_{k\le m} \|\mathbf{X}\boldsymbol{\psi}_k\|^2,
\end{aligned}$$

because $\boldsymbol{x}_i^\intercal \boldsymbol{\psi}_k$s are the components of the vector $\mathbf{X}\boldsymbol{\psi}_k$. Therefore to minimize the distance of the observed data to the subspace, i.e. $\sum_{i\le n} \|\boldsymbol{x}_i - \boldsymbol{w}_i\|^2$, we need to maximize $\sum_{k\le m} \|\mathbf{X}\boldsymbol{\psi}_k\|^2$. Hence to show that V is the closest m-dimensional linear subspace to the data we have to show that

$$\sum_{k\le m} \|\mathbf{X}\boldsymbol{\psi}_k\|^2 \le \sum_{k\le m} \|\mathbf{z}_k\|^2 = \sum_{k\le m} \lambda_k.$$

When $k = 1$ this is easy, since we have

$$\|\mathbf{X}\boldsymbol{\psi}_1\|^2 = \boldsymbol{\psi}_1^\intercal \mathbf{X}^\intercal \mathbf{X}\boldsymbol{\psi}_1 \le \max_{\|\boldsymbol{\phi}\|=1} \boldsymbol{\phi}^\intercal \mathbf{X}^\intercal \mathbf{X}\boldsymbol{\phi} = \lambda_1.$$

Or equivalently we can say that $\mathbf{z}_1$ has the largest variance among all linear combinations $\mathbf{X}\boldsymbol{\phi}$ with $\|\boldsymbol{\phi}\| = 1$. For general k note that we have

$$\boldsymbol{\psi}_k = (\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_1)\boldsymbol{\phi}_1 + \cdots + (\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_p)\boldsymbol{\phi}_p,$$

since $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p$ is an orthonormal basis for $\mathbb{R}^p$. Hence

$$\begin{aligned}
\|\mathbf{X}\boldsymbol{\psi}_k\|^2 = \boldsymbol{\psi}_k^\intercal \mathbf{X}^\intercal \mathbf{X}\boldsymbol{\psi}_k &= \boldsymbol{\psi}_k^\intercal \mathbf{X}^\intercal \mathbf{X}\big((\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_1)\boldsymbol{\phi}_1 + \cdots + (\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_p)\boldsymbol{\phi}_p\big) \\
&= \boldsymbol{\psi}_k^\intercal \big((\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_1)\mathbf{X}^\intercal \mathbf{X}\boldsymbol{\phi}_1 + \cdots + (\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_p)\mathbf{X}^\intercal \mathbf{X}\boldsymbol{\phi}_p\big) \\
&= \boldsymbol{\psi}_k^\intercal \big((\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_1)\lambda_1\boldsymbol{\phi}_1 + \cdots + (\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_p)\lambda_p\boldsymbol{\phi}_p\big) \\
&= \lambda_1(\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_1)^2 + \cdots + \lambda_p(\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_p)^2.
\end{aligned}$$

Thus we get

$$\sum_{k\le m} \|\mathbf{X}\boldsymbol{\psi}_k\|^2 = \sum_{k\le m}\sum_{l\le p} \lambda_l(\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_l)^2 = \sum_{l\le p} \lambda_l \sum_{k\le m} (\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_l)^2 = \sum_{l\le p} a_l\lambda_l,$$

where $a_l := \sum_{k \leq m} (\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_l)^2 \geq 0$. We also know that

$$a_l = \sum_{k \leq m} (\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_l)^2 = \sum_{k \leq m} (\boldsymbol{\phi}_l^\intercal \boldsymbol{\psi}_k)^2 \leq \|\boldsymbol{\phi}_l\|^2 = 1,$$

since $\boldsymbol{\phi}_l^\intercal \boldsymbol{\psi}_k$s are the coefficients of the representation of $\boldsymbol{\phi}_l$ in an orthonormal basis containing $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_m$. In addition note that

$$\sum_{l \leq p} a_l = \sum_{l \leq p} \sum_{k \leq m} (\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_l)^2 = \sum_{k \leq m} \sum_{l \leq p} (\boldsymbol{\psi}_k^\intercal \boldsymbol{\phi}_l)^2 = \sum_{k \leq m} \|\boldsymbol{\psi}_k\|^2 = \sum_{k \leq m} 1 = m.$$

Therefore $\sum_{l \leq p} a_l \lambda_l$ is at most $\sum_{l \leq m} \lambda_l$, since λ_ls are ordered in a decreasing way. Hence we have shown that $\sum_{k \leq m} \|\mathbf{X}\boldsymbol{\psi}_k\|^2 \leq \sum_{l \leq m} \lambda_l$, as desired.

Chapter 5

Nonlinear Methods

5.1 Splines

We have seen in Section 2.3.1 that we can apply nonlinear basis functions, like polynomials, to the features to obtain new features. Then we can use these new features in linear regression. These new features can also be used in other learning methods such as logistic regression and LDA. One particularly useful set of basis functions is the set of piecewise polynomial functions. For example, using a feature X we can construct the features

$$\begin{aligned} a_0(X) &= I(X < c), \\ a_1(X) &= I(X < c)X, \\ &\vdots \\ a_m(X) &= I(X < c)X^m, \\ b_0(X) &= I(X \geq c), \\ b_1(X) &= I(X \geq c)X, \\ &\vdots \\ b_m(X) &= I(X \geq c)X^m, \end{aligned}$$

where I is the indicator function. A linear combination of a_j, b_j is a piecewise degree m polynomial of X whose coefficients can change at the **knot** c.

Using these new nonlinear functions of X can result in much more flexible models. However a drawback of using arbitrary piecewise polynomials is that they can be discontinuous. And we would like to obtain the coefficients of a_j, b_j in a way that the resulting piecewise polynomial function, and possibly some of its derivatives, are continuous everywhere, especially at c. But we find the coefficients of a_j, b_j by least squares method, and it seems hard to impose the additional continuity

constraints while we are fitting the model. The idea for avoiding this difficulty is to impose the constraints on the basis functions. Then we can use those new basis functions which satisfy the constraints to fit the model. Let us explain this point further.

Suppose we want to model the relationship between Y, X by a degree m piecewise polynomial which has continuous derivatives up to order $m-1$, and has knots at $c_1, \ldots, c_k$. Then we can write

$$\begin{aligned} Y = \beta_0 + \beta_1 X + \cdots + \beta_m X^m \\ + \beta_{m+1}(X - c_1)_+^m + \cdots + \beta_{m+k}(X - c_k)_+^m + \epsilon, \end{aligned}$$

where

$$(X - c_j)_+^m = \begin{cases} 0 & X < c_j, \\ (X - c_j)^m & X \geq c_j. \end{cases}$$

Note that each of the basis functions $1, x, x^2, \ldots, x, (x - c_j)_+^m$ are (piecewise) polynomials of degree at most m, and they all have everywhere continuous derivatives up to order $m - 1$. Hence the function

$$\beta_0 + \beta_1 X + \cdots + \beta_m X^m + \beta_{m+1}(X - c_1)_+^m + \cdots + \beta_{m+k}(X - c_k)_+^m$$

is a piecewise degree m polynomial of X, which automatically has continuous derivatives up to order $m-1$. Functions of this form are called degree m **splines**. Now we can estimate the coefficients β_j by our learning method, for example least squares, and there is no need to impose the constraints during estimation, since the constraints are automatically satisfied.

We can also use the above idea to impose more constraints, which are sometimes useful. For example, in **natural cubic splines**, in which $m = 3$, we also require the function to be linear in the first and last intervals, i.e. $(-\infty, c_1)$ and (c_k, ∞). Consider a cubic spline

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - c_1)_+^3 + \cdots + \beta_{k+3}(x - c_k)_+^3.$$

When $x < c_1$ we have $(x - c_j)_+^3 = 0$. So the linearity constraint on $(-\infty, c_1)$ implies that $\beta_2 = \beta_3 = 0$. The linearity constraint on (c_k, ∞) will yield relationships between $\beta_4, \ldots, \beta_{k+3}$, which after simplifications imply that

$$f(x) = \alpha_1 + \alpha_2 x + \alpha_3 N_3(x) + \cdots + \alpha_k N_k(x),$$

where

$$N_{j+2}(x) = \frac{(x - c_j)_+^3 - (x - c_k)_+^3}{c_k - c_j} - \frac{(x - c_{k-1})_+^3 - (x - c_k)_+^3}{c_k - c_{k-1}}.$$

To simplify the notation we also use $N_1(x) = 1$ and $N_2(x) = x$. Although the computations for deriving the above relation is cumbersome, but it can readily be

seen that each N_j is a piecewise polynomial of degree at most 3 with continuous derivatives up to order 2, which is also linear on the intervals $(-\infty, c_1)$ and (c_k, ∞). Because for $x < c_1$ we have $(x - c_j)_+^3 = 0$ for each j, and for $x > c_k$ we have

$$(x - c_j)_+^3 = (x - c_j)^3 = x^3 - 3c_j x^2 + 3c_j^2 x - c_j^3;$$

so the coefficients of x^3, x^2 vanish in $N_j(x)$.

5.2 Smoothing Splines

Suppose we have a feature X with observations $x_1, \ldots, x_n$, and a response $Y = f(X) + \epsilon$ with corresponding observations $y_1, \ldots, y_n$. In order to estimate f we have so far assumed that f has a given form with some parameters; then we estimated those parameters. For example, we may assume that f is a linear spline with knots at $c_1, \ldots, c_k$, i.e. a linear combination of $1, x, (x - c_1)_+, \ldots, (x - c_k)_+$, and then we can estimate the coefficients of that linear combination by minimizing the residual sum of squares.

Alternatively, we can look for the estimate of f in a large class of functions by a different minimization method. Consider the functional

$$F[g] = \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \big(g''(x)\big)^2 dx,$$

which is defined for a sufficiently smooth function g. (The integral can be taken over $(-\infty, \infty)$ or a bounded interval containing the range of values of X, but its bounds are not important in what follows.) Also, λ is a fixed tuning parameter. We will show that the functional F has a minimizer, which is know as a **smoothing spline**.

In fact, we will show that the functional F achieves its minimum in the space of natural cubic splines with knots at $x_1, \ldots, x_n$. Hence smoothing splines are natural cubic splines; however, their coefficients are different from the coefficients of a natural cubic spline which minimizes RSS. Let g be an arbitrary function in the domain of the functional. We will show that there is a natural cubic spline h such that $F[h] \leq F[g]$. Hence if F has a minimizer, it must be a natural cubic spline. Then we show that F actually achieves its minimum in the space of natural cubic splines, as desired.

Let $N_1, \ldots, N_n$ be the basis for the space of natural cubic splines with knots at $x_1, \ldots, x_n$. We assume that $x_1, \ldots, x_n$ are distinct points. To simplify the notation we also assume that $x_1, \ldots, x_n$ are ordered in an increasing way. Let $\mathbf{N} = [N_j(x_i)]$ be the $n \times n$ matrix resulting from applying this basis to the observed predictor

values. Remember that $N_1(x) = 1$ and $N_2(x) = x$. The other entries of $\mathbf{N}$ are

$$\begin{aligned} N_{j+2}(x_i) &= \frac{(x_i - x_j)_+^3 - (x_i - x_n)_+^3}{x_n - x_j} - \frac{(x_i - x_{n-1})_+^3 - (x_i - x_n)_+^3}{x_n - x_{n-1}} \\ &= \frac{(x_i - x_j)_+^3}{x_n - x_j} - \frac{(x_i - x_{n-1})_+^3}{x_n - x_{n-1}} \qquad \text{(since } x_i \le x_n\text{)} \\ &= \begin{cases} 0 & \text{if } i \le j, \\ \frac{(x_i - x_j)^3}{x_n - x_j} & \text{if } j < i < n, \\ (x_n - x_j)^2 - (x_n - x_{n-1})^2 & \text{if } i = n. \end{cases} \end{aligned}$$

Note that we have $N_{j+2}(x_i) \neq 0$ for $i > j$. The matrix $\mathbf{N}$ is almost lower triangular, and it can be easily seen that it has at least $n - 1$ linearly independent columns. But to show that $\mathbf{N}$ is invertible we have to prove that all its columns are linearly independent.

Suppose there is a linear dependence relation between the columns of $\mathbf{N}$. Then there are coefficients $a_1, \ldots, a_n$ such that

$$a_1 N_1(x_i) + \cdots + a_n N_n(x_i) = 0$$

for every $i \le n$. Let $\tilde{h}(x) := a_1 N_1(x) + \cdots + a_n N_n(x)$. Then $\tilde{h}$ is a piecewise cubic polynomial function whose first and second derivatives are continuous. Also, for every $i \le n$ we have $\tilde{h}(x_i) = 0$; so $\tilde{h}$ has n distinct roots. But this implies that $\tilde{h}'$ has $n - 1$ distinct roots, one between each pair x_i, x_{i+1}, due to the mean value theorem. And therefore $\tilde{h}''$ has $n - 2$ distinct roots, one between each pair of consecutive roots of $\tilde{h}'$.

However, note that $\tilde{h}''$ is a linear function in each interval $[x_i, x_{i+1}]$. In addition, $\tilde{h}'' = 0$ on the intervals $(-\infty, x_1]$ and $[x_n, \infty)$, since $\tilde{h}$ is linear on these intervals. Suppose $\tilde{h}''$ is not identically zero on $[x_k, x_l]$, and it vanishes over some intervals before x_k and after x_l. Then $\tilde{h}''$ cannot have any roots in $(x_k, x_{k+1}]$ and $[x_{l-1}, x_l)$, since over these intervals $\tilde{h}''$ is a nonzero linear function that vanishes at one of the endpoints. Furthermore, $\tilde{h}''$ has at most one root in each of the $l - k - 2$ intervals

$$(x_{k+1}, x_{k+2}], (x_{k+2}, x_{k+3}], \ldots, (x_{l-2}, x_{l-1}),$$

since $\tilde{h}''$ is a nonzero linear function over each of these intervals. But by our assumption, between the $l - k + 1$ roots $x_k, x_{k+1}, \ldots, x_l$ of $\tilde{h}$, $\tilde{h}''$ must have at least $l - k - 1$ distinct roots. This contradiction implies that $\tilde{h}''$ must be identically zero everywhere, which implies that the coefficients $a_1, \ldots, a_n$ are all zero. So the columns of $\mathbf{N}$ are linearly independent; and hence $\mathbf{N}$ is an invertible matrix.

Now let $\mathbf{g} = \begin{bmatrix} g(x_1) & \ldots & g(x_n) \end{bmatrix}^\intercal$. Then there is a vector $\boldsymbol{\beta}$ such that $\mathbf{N}\boldsymbol{\beta} = \mathbf{g}$. Let

$$h(x) = \beta_1 N_1(x) + \cdots + \beta_n N_n(x).$$

Then h is a natural cubic spline that satisfies $h(x_i) = g(x_i)$ for every i. Therefore $\sum_{i=1}^n (y_i - g(x_i))^2 = \sum_{i=1}^n (y_i - h(x_i))^2$. Hence we have

$$F[g] - F[h] = \lambda \int (g'')^2 - (h'')^2 dx.$$

Thus to conclude $F[h] \leq F[g]$ we only need to show that $\int (h'')^2 dx \leq \int (g'')^2 dx$. To prove we integrate by parts twice to obtain (Note that we have to decompose the domain of integration before integrating by parts, because $h''', h^{(4)}$ do not exist everywhere.)

$$\begin{aligned}
\int (g'' - h'')h'' dx &= \int_{x_1}^{x_n} (g'' - h'')h'' dx \qquad \text{(since } h'' = 0 \text{ outside } [x_1, x_n]) \\
&= \sum_{i=2}^{n} \int_{x_{i-1}}^{x_i} (g'' - h'')h'' dx \\
&= -\sum_{i=2}^{n} \int_{x_{i-1}}^{x_i} (g' - h')h''' dx + \sum_{i=2}^{n} \big[(g'(x_i) - h'(x_i))h''(x_i) \\
&\qquad\qquad - (g'(x_{i-1}) - h'(x_{i-1}))h''(x_{i-1})\big] \\
&= \sum_{i=2}^{n} \int_{x_{i-1}}^{x_i} (g - h)h^{(4)} dx - \sum_{i=2}^{n} \big[(g(x_i) - h(x_i))h'''(x_i) \\
&\qquad\qquad - (g(x_{i-1}) - h(x_{i-1}))h'''(x_{i-1})\big] \\
&\qquad + \sum_{i=1}^{n} \big[(g'(x_i) - h'(x_i))h''(x_i) \\
&\qquad\qquad - (g'(x_{i-1}) - h'(x_{i-1}))h''(x_{i-1})\big].
\end{aligned}$$

Now note that the integrals containing $h^{(4)}$ vanish since h is a piecewise cubic polynomial. We also have $g(x_i) = h(x_i)$ for every i; so the n-term sum in the above formula which contains h''' vanishes too. In addition, the n-term sum containing h'' is a telescoping sum; hence it equals $(g'(x_n) - h'(x_n))h''(x_n) - (g'(x_1) - h'(x_1))h''(x_1)$. But $h''(x_n) = h''(x_1) = 0$, because h is a natural spline, so it is linear on the intervals $(-\infty, x_1)$ and (x_n, ∞). Therefore all the terms in the above equation vanish. Thus we have $\int (g'' - h'')h'' dx = 0$. Hence we get

$$\int (h'')^2 dx = \int h''g'' dx \leq \int \frac{1}{2}((h'')^2 + (g'')^2) dx = \frac{1}{2}\int (h'')^2 dx + \frac{1}{2}\int (g'')^2 dx.$$

Therefore we obtain $\int (h'')^2 dx \leq \int (g'')^2 dx$, as desired.

Now let us compute $F[h]$. We have

$$\begin{aligned} F[h] &= \sum_{i=1}^{n}(y_i - h(x_i))^2 + \lambda \int \big(h''(x)\big)^2 dx \\ &= \sum_{i=1}^{n}\big(y_i - \sum_{j=1}^{n} N_j(x_i)\beta_j\big)^2 + \lambda \int \big(\sum_{j=1}^{n} N_j''(x)\beta_j\big)^2 dx \\ &= \sum_{i=1}^{n}\big(y_i - \sum_{j=1}^{n} N_j(x_i)\beta_j\big)^2 + \lambda \sum_{j,k}\beta_k\beta_j \int N_j''(x)N_k''(x)dx \\ &= \|\mathbf{y} - \mathbf{N}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\intercal\boldsymbol{\Omega}\boldsymbol{\beta}, \end{aligned}$$

where $\boldsymbol{\Omega}$ is the $n \times n$ matrix with entries $\boldsymbol{\Omega}_{jk} = \int N_j''(x)N_k''(x)dx$. Therefore to minimize F over the space of natural cubic splines we need to find $\boldsymbol{\beta}$ that minimizes $\|\mathbf{y} - \mathbf{N}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\intercal\boldsymbol{\Omega}\boldsymbol{\beta}$. We can differentiate this expression with respect to $\boldsymbol{\beta}$, and similarly to the computations in Section 4.4.1 for ridge regression, we obtain that the minimizer $\hat{\boldsymbol{\beta}}$ satisfies

$$(\mathbf{N}^\intercal\mathbf{N} + \lambda\boldsymbol{\Omega})\hat{\boldsymbol{\beta}} = \mathbf{N}^\intercal\mathbf{y}.$$

And when the matrix $\mathbf{N}^\intercal\mathbf{N} + \lambda\boldsymbol{\Omega}$ is invertible, the vector

$$\hat{\boldsymbol{\beta}} = (\mathbf{N}^\intercal\mathbf{N} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{N}^\intercal\mathbf{y}$$

provides the minimizer $\hat{h}(x) = \hat{\beta}_1 N_1(x) + \cdots + \hat{\beta}_n N_n(x)$ for the functional F.

5.3 Local Regression

Suppose we have a feature X with observations $x_1, \ldots, x_n$. In order to estimate $f(x)$ using **local regression** we first need to choose a **kernel** $K(x_i, x)$ which assigns a weight to the observation x_i with respect to x. The kernel is chosen so that it assigns a low (or zero) weight to an observation x_i which is far from x, and a higher weight to observations close to x. Usually there is a parameter in the kernel K that controls the width of the region in which K assigns a higher weight, and this parameter should be fixed beforehand. Then we use these weights in weighted least squares method, described in Section 2.3.3, to fit the model by minimizing

$$\sum_{i=1}^{n} w_i(x)(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} w_i(x)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \cdots - \hat{\beta}_m x_i^m)^2,$$

where $w_i(x) = K(x_i, x)$ are the weights. As you see, we can use linear or polynomial functions in this method. Note that the weights depend on x; so the above fitting procedure will only give us $\hat{f}(x)$ for this particular value of x. And if we want to estimate $\hat{f}(x_0)$ for another value x_0 we need to compute the weights $w(x_0) = K(x_i, x_0)$, and fit the weighted least squares method again.