Chapter 3

Multiple Linear Regression

The general purpose of multiple linear regression is to seek for the linear relationship between a dependent variable and several independent variables. Multiple regression allows researchers to examine the effect of more than one independent variables on response at the same time. For some research questions, regression can be used to examine how much a particular set of independent variables can explain sufficiently the outcome. In other cases, multiple regression is used to examine the effect of outcome while accounting for more than one factor that could influence the outcome. In this chapter we discuss multiple linear regression. To facilitate the discussion of the theory of the multiple regression model we start with a brief introduction of the linear space and the projection in the linear space. Then we will introduce multiple linear model in matrix form. All subsequent discussions of multiple regression will be based on its matrix form.

3.1 Vector Space and Projection

First we briefly discuss the vector space, subspace, projection, and quadratic form of multivariate normal variable, which are useful in the discussions of the subsequent sections of this chapter.

3.1.1 *Vector Space*

A vector is a geometric object which has both magnitude and direction. A vector is frequently represented by a line segment connecting the initial point A with the terminal point B and denoted by $\overrightarrow{AB}$. The magnitude of the vector $\overrightarrow{AB}$ is the length of the segment and the direction of this vector characterizes the displacement of the point B relative to the point

A. Vectors can be added, subtracted, multiplied by a number, and flipped around (multiplying by number -1) so that the direction is reversed. These operations obey the familiar algebraic laws: commutativity, associativity, and distributivity. The sum of two vectors with the same initial point can be found geometrically using the parallelogram law. Multiplication by a positive number, commonly called a scalar in this context, amounts to changing the magnitude of vector, that is, stretching or compressing it while keeping its direction; multiplication by -1 preserves the magnitude of the vector but reverses its direction. Cartesian coordinates provide a systematic way of describing vectors and operations on them.

A vector space is a set of vectors that is closed under finite vector addition and scalar multiplication. The basic example is n-dimensional Euclidean space, where every element is represented by a list of real numbers, such as

$$\boldsymbol{x}^{'} = (x_1, x_2, \cdots, x_n).$$

Scalars are real numbers, addition is componentwise, and scalar multiplication is multiplication on each term separately. Suppose V is closed under vector addition on $\mathbb{R}^n$: if $u, v \in V$, then $u + v \in V$. V is also closed under scalar multiplication: if $a \in \mathbb{R}^1$, $v \in V$, then $av \in V$. Then V is a vector space (on $\mathbb{R}^n$). We will focus our discussion only on vector space on n-dimensional Euclidean space. For example, for any positive integer n, the space of all n-tuples of elements of real line $\mathbb{R}^1$ forms an n-dimensional real vector space sometimes called real coordinate space and denoted by $\mathbb{R}^n$. An element in $\mathbb{R}^n$ can be written as

$$\boldsymbol{x}^{'} = (x_1, x_2, \cdots, x_n),$$

where each x_i is an element of $\mathbb{R}^1$. The addition on $\mathbb{R}^n$ is defined by

$$\boldsymbol{x} + \boldsymbol{y} = (x_1 + y_1,\ x_2 + y_2,\ \cdots,\ x_n + y_n),$$

and the scalar multiplication on $\mathbb{R}^n$ is defined by

$$a\,\boldsymbol{x} = (ax_1,\ ax_2,\ \cdots,\ ax_n).$$

When $a = -1$ the vector $a\boldsymbol{x}$ has the same length as $\boldsymbol{x}$ but with a geometrically reversed direction.

It was F. Hausdorff who first proved that every vector space has a basis. A basis makes it possible to express every vector of the space as a unique tuple of the field elements, although caution must be exercised when a vector space does not have a finite basis. In linear algebra, a basis is a set of vectors that, in a linear combination, can represent every vector in a given vector space, and such that no element of the set can be represented as a linear combination of the others. In other words, a basis is a linearly independent spanning set. The following is an example of basis of $\mathbb{R}^n$:

$$\begin{aligned}
\boldsymbol{e}_1^{'} &= (1,\ 0,\ 0, \cdots,\ 0)_{1\times n}\\
\boldsymbol{e}_2^{'} &= (0,\ 1,\ 0, \cdots,\ 0)_{1\times n}\\
\boldsymbol{e}_3^{'} &= (0,\ 0,\ 1, \cdots,\ 0)_{1\times n}\\
&\vdots\\
\boldsymbol{e}_n^{'} &= (0,\ 0,\ 0, \cdots,\ 1)_{1\times n}.
\end{aligned}$$

Actually, the above vectors consist of the standard orthogonal basis of the vector space $\mathbb{R}^n$. Any vector $\boldsymbol{x}^{'} = (x_1, x_2, \cdots, x_n)$ in the $\mathbb{R}^n$ can be a linear combination of $\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n$. In fact,

$$\boldsymbol{x} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 + x_3\boldsymbol{e}_3 + \cdots + x_n\boldsymbol{e}_n.$$

This representation is unique. i.e., if there is another representation such that

$$\boldsymbol{x} = x_1^*\boldsymbol{e}_1 + x_2^*\boldsymbol{e}_2 + x_3^*\boldsymbol{e}_3 + \cdots + x_n^*\boldsymbol{e}_n,$$

then

$$\begin{aligned}
&(x_1 - x_1^*)\boldsymbol{e}_1 + (x_2 - x_2^*)\boldsymbol{e}_2 + \cdots + (x_n - x_n^*)\boldsymbol{e}_n\\
&= (x_1 - x_1^*,\ x_2 - x_2^*,\ \cdots, x_2 - x_2^*) = (0,\ 0, \cdots,\ 0).
\end{aligned}$$

Therefore, we have $x_i = x_i^*$ for all $i = 1, 2, \cdots, n$.

Given a vector space V, a nonempty subset W of V that is closed under addition and scalar multiplication is called a subspace of V. The intersection of all subspaces containing a given set of vectors is called its span. If no vector can be removed without changing the span, the vectors in this set is said to be linearly independent. A linearly independent set whose span is V is called a basis for V. A vector span by two vectors v and w can be defined as: $\boldsymbol{x} : \boldsymbol{x} = av + bw,\ \ \text{for all}\ \ (a, b) \in \mathbb{R}^2$. Note that v and w may not be necessarily independent. If a vector space S is spanned by a set of independent vectors $v_1, v_2, \cdots, v_p$, i.e., S is the set of vectors

$$\{\boldsymbol{x} : \boldsymbol{x} = a_1 + v_1 + a_2v_2 + \cdots + a_pv_p,\ \ \text{for all}\ \ (a_1, a_2, \cdots, a_p) \in \mathbb{R}^p\},$$

then the dimension of S is p. Vectors $v_1, v_2, \cdots, v_p$ are the basis of the vector space S. The dimension of a vector space S is the largest number of a set of independent vectors in S. If the dimension of a linear space S is p we write $\text{Dim}(S) = p$.

3.1.2 *Linearly Independent Vectors*

If there exist a finite number of distinct vectors $v_1, v_2, \cdots, v_n$ in vector space V and scalars $a_1, a_2, \cdots, a_n$, not all zero, such that

$$a_1 v_1 + a_2 v_2 + a_3 v_3 + \cdots + a_n v_n = 0,$$

then the vectors $v_1, v_2, \cdots, v_n$ are said to be linearly dependent. If $v_1, v_2, \cdots, v_n$ are dependent then out of these n vectors there is at least one vector that can be expressed as a linear combination of other vectors. Note that the zero on the right is the zero vector, not the number zero. If no such scalars exist, then the vectors $v_1, v_2, \cdots, v_n$ are said to be linearly independent. This condition can be reformulated as follows: whenever $a_1, a_2, \cdots, a_n$ are scalars such that

$$a_1 v_1 + a_2 v_2 + a_3 v_3 + \cdots + a_n v_n = 0,$$

we have $a_i = 0$ for $i = 1, 2, \cdots, n$, then $v_1, v_2, \cdots, v_n$ are linearly independent.

A basis of a vector space V is defined as a subset of vectors in V that are linearly independent and these vectors span space V. Consequently, if $(v_1, v_2, \cdots, v_n)$ is a list of vectors in V, then these vectors form a basis if and only if every vector $\boldsymbol{x} \in V$ can be uniquely expressed by a linear combination of $v_1, v_2, \cdots, v_p$. i.e.,

$$\boldsymbol{x} = a_1 v_1 + a_2 v_2 + \cdots + a_n v_n, \text{ for any } \boldsymbol{x} \in \text{V}.$$

The number of basis vectors in V is called the dimension of linear space V. Note that a vector space can have more than one basis, but the number of vectors which form a basis of the vector space V is always fixed. i.e., the dimension of vector space V is fixed but there will be more than one basis. In fact, if the dimension of vector space V is n, then any n linearly independent vectors in V form its basis.

3.1.3 *Dot Product and Projection*

If $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ are two vectors in a vector Euclidean space $\mathbb{R}^n$. The dot product of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$\boldsymbol{x} \cdot \boldsymbol{y} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

Two vectors are said to be orthogonal if their dot product is 0. If θ is the angle between two vectors $(x_1, x_2, \cdots, x_n)$ and $(y_1, y_2, \cdots, y_n)$, the cosine of the angle between the two vectors is defined as

$$\cos(\theta) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|} = \frac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}\sqrt{y_1^2 + y_2^2 + \cdots + y_n^2}} \tag{3.1}$$

Two orthogonal vectors meet at 90°; i.e., they are perpendicular. One important application of the dot product is projection. The projection of a vector $\boldsymbol{y}$ onto another vector $\boldsymbol{x}$ forms a new vector that has the same direction as the vector $\boldsymbol{x}$ and the length $|\boldsymbol{y}|\cos(\theta)$, where $|\boldsymbol{y}|$ denotes the length of vector $\boldsymbol{y}$ and θ is the angle between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. We write this projection as $P_{\boldsymbol{x}}\boldsymbol{y}$. The projection vector can be expressed as

$$\begin{aligned} P_{\boldsymbol{x}}\boldsymbol{y} &= |\boldsymbol{y}|\cos(\theta)\frac{\boldsymbol{x}}{|\boldsymbol{x}|} = |\boldsymbol{y}|\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}\frac{\boldsymbol{x}}{|\boldsymbol{x}|} \\ &= \frac{x_1y_1 + x_2y_2 + \cdots + x_ny_n}{x_1^2 + x_2^2 + \cdots + x_n^2}\boldsymbol{x} = \lambda\boldsymbol{x}, \end{aligned} \tag{3.2}$$

where λ is a scalar and

$$\lambda = \frac{x_1y_1 + x_2y_2 + \cdots + x_ny_n}{x_1^2 + x_2^2 + \cdots + x_n^2} = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\boldsymbol{x}\boldsymbol{x}}.$$

Thus, the projection of $\boldsymbol{y}$ onto vector $\boldsymbol{x}$ is a vector $\boldsymbol{x}$ multiplying a scalar λ where λ is the $\cos(\theta)$ and θ is the angle between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

If $\boldsymbol{x}$ and $\boldsymbol{y}$ are two vectors in $\mathbb{R}^n$. Consider the difference vector between the vector $\boldsymbol{e}$, $\boldsymbol{e} = \lambda\boldsymbol{x} - \boldsymbol{y}$, and $\lambda = \boldsymbol{x}\cdot\boldsymbol{y}/\boldsymbol{x}\cdot\boldsymbol{x}$. The vector $\boldsymbol{e}$ is perpendicular to the vector $\boldsymbol{x}$ when $\lambda = (\boldsymbol{x}\cdot\boldsymbol{y})/(\boldsymbol{x}\cdot\boldsymbol{x})$. To see this we simply calculate the dot product of $\boldsymbol{e}$ and $\boldsymbol{x}$:

$$\boldsymbol{e}\cdot\boldsymbol{x} = (\lambda\boldsymbol{x} - \boldsymbol{y})\cdot\boldsymbol{x} = \lambda\boldsymbol{x}\cdot\boldsymbol{x} - \boldsymbol{x}\cdot\boldsymbol{y} = \Big(\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\boldsymbol{x}\cdot\boldsymbol{x}}\Big)\boldsymbol{x}\cdot\boldsymbol{x} - \boldsymbol{x}\cdot\boldsymbol{y} = 0$$

Thus, the angle between $\boldsymbol{e}$ and $\boldsymbol{x}$ is indeed 90^o, i.e., they are perpendicular to each other. In addition, since $\boldsymbol{e}$ is perpendicular to $\boldsymbol{x}$, it is the vector with the shortest distance among all the vectors starting from the end of $\boldsymbol{y}$ and ending at any point on $\boldsymbol{x}$.

If a vector space has a basis and the length of the basis vectors is a unity then this basis is an orthonormal basis. Any basis divided by its length forms an orthonormal basis. If S is a p-dimensional subspace of a vector space V, then it is possible to project vectors in V onto S. If the subspace S has an orthonormal basis $(w_1, w_2, \cdots, w_p)$, for any vector $\boldsymbol{y}$ in V, the projection of $\boldsymbol{y}$ onto the subspace S is

$$P_S\boldsymbol{y} = \sum_{i=1}^{p}(\boldsymbol{y}\cdot w_i)w_i. \tag{3.3}$$

Let vector spaces S and T be the two subspaces of a vector space V and union $S \cup T = V$. If for any vector $\boldsymbol{x} \in S$ and any vector $\boldsymbol{y} \in T$, the dot product $\boldsymbol{x}\cdot\boldsymbol{y} = 0$, then the two vector spaces S and T are said to be orthogonal. Or we can say that T is the orthogonal space of S, denoted by

$T = S^{\perp}$. Thus, for a vector space V, if S is a vector subspace in V, then $V = S \cup S^{\perp}$. Any vector $\boldsymbol{y}$ in V can be written uniquely as $\boldsymbol{y}_S + \boldsymbol{y}_S^{\perp}$, where $\boldsymbol{y}_S \in S$ and $\boldsymbol{y}_S^{\perp}$ is in $S^{\perp}$, the orthogonal subspace of S.

A projection of a vector onto a linear space S is actually a linear transformation of the vector and can be represented by a projection matrix times the vector. A projection matrix P is an $n \times n$ square matrix that gives the projection from $\mathbb{R}^n$ onto subspace S. The columns of P are the projections of the standard basis vectors, and S is the image of P. For the projection matrix we have the following theorems.

Theorem 3.1. *A square matrix P is a projection matrix if and only if it is idempotent, i.e., $P^2 = P$.*

Theorem 3.2. *Let $U = (u_1, u_2, \cdots, u_k)$ be an orthonormal basis for a subspace W of linear space V. The matrix $UU^{'}$ is a projection matrix of V onto W. i.e., for any vector $v \in V$ the projection of v onto W is $Proj_W v = UU^{'}v$.*

The matrix $UU^{'}$ is called the projection matrix for the subspace W. It does not depend on the choice of orthonormal basis. If we do not start with an orthonormal basis of W, we can still construct the projection matrix. This can be summarized in the following theorem.

Theorem 3.3. *Let $A = (a_1, a_2, \cdots, a_k)$ be any basis for a subspace W of V. The matrix $A(A^{'}A)^{-1}A^{'}$ is a projection matrix of V onto W. i.e., for any vector $v \in V$ the projection of v onto W is*

$$Proj_W v = A(A^{'}A)^{-1}A^{'}v. \tag{3.4}$$

To understand the above three theorems the following lemma is important.

Lemma 3.1. *Suppose that A is an $n \times k$ matrix whose columns are linearly independent. Then $AA^{'}$ is invertible.*

Proof. Consider the transformation A: $\mathbb{R}^k \to \mathbb{R}^k$ determined by A. Since the columns of A are linearly independent, this transformation is one-to-one. In addition, the null space of $A^{'}$ is orthogonal to the column space of A. Thus, $A^{'}$ is one-to-one on the column space of A, and as a result, $A^{'}A$ is one-to-one transformation $\mathbb{R}^k \to \mathbb{R}^k$. By invertible matrix theorem, $A^{'}A$ is invertible. □

Let's now derive the projection matrix for the column space of A. Note that any element of the column space of A is a linear combination of the columns of A, i.e., $x_1a_1 + x_2a_2 + \cdots + x_ka_k$. If we write

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix},$$

then we have

$$x_1a_1 + x_2a_2 + \cdots + x_ka_k = A\boldsymbol{x}.$$

Now, for any vector $v \in \mathbb{R}^n$, we denote the projection of v onto W by x_p.

$$Proj_Wv = Ax_p.$$

The projection matrix can be found by calculating x_p. The projection of vector v onto W is characterized by the fact that $v - Proj_Wv$ is orthogonal to any vector w in W. Thus we have

$$w \cdot (v - Proj_Wv) = 0$$

for all w in W. Since $w = A\boldsymbol{x}$ for some $\boldsymbol{x}$, we have

$$A\boldsymbol{x} \cdot (v - Ax_p) = 0$$

for all $\boldsymbol{x}$ in $\mathbb{R}^n$. Write this dot product in terms of matrices yields

$$(A\boldsymbol{x})^{'}(v - A\boldsymbol{x}_p) = 0$$

which is equivalent to

$$(x^{'}A^{'})(v - A\boldsymbol{x}_p) = 0$$

Converting back to dot products we have

$$\boldsymbol{x} \cdot A^{'}(v - A\boldsymbol{x}_p) = 0$$

We get

$$A^{'}v = A^{'}A\boldsymbol{x}_p$$

Since $A^{'}A$ is invertible we have

$$(A^{'}A)^{-1}A^{'}v = \boldsymbol{x}_p$$

Since $A\boldsymbol{x}_p$ is the desired projection, we have

$$A(A^{'}A)^{-1}A^{'}v = A\boldsymbol{x}_p = Proj_Wv$$

Therefore, we conclude that the projection matrix for W is $A(A^{'}A)^{-1}A^{'}$.

Projection matrix is very useful in the subsequent discussions of linear regression model $Y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. A squared matrix, $P = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^{'})^{-1}\boldsymbol{X}^{'}$, is constructed using the design matrix. It can be easily verified that P is an idempotent matrix:

$$P^2 = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^{'})^{-1}\boldsymbol{X}^{'}\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^{'})^{-1}\boldsymbol{X}^{'} = P.$$

Thus, $P = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^{'})^{-1}\boldsymbol{X}^{'}$ is a projection matrix. In addition, if we define a matrix as $I - P = I - \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^{'})^{-1}\boldsymbol{X}^{'}$. It is easy to see that $I - P$ is also idempotent. In fact,

$$(I - P)^2 = I - 2P + P^2 = I - 2P + P = I - P.$$

Therefore, $I - P = I - \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X}^{'})^{-1}\boldsymbol{X}^{'}$ is a projection matrix. In the subsequent sections we will see how these projection matrices are used to obtain the best linear unbiased estimator (BLUE) for the linear regression model and how they are used in regression model diagnosis.

3.2 Matrix Form of Multiple Linear Regression

In many scientific research it is often needed to determine the relationship between a response (or dependent) variable (y) and more than one regressors (or independent variables) ($x_1, x_2, \cdots, x_k$). A general form of a multiple linear regression model is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{3.5}$$

where ε is the random error. Here, regressors $x_1, x_2, \cdots, x_k$ may contain regressors and their higher order terms. In the classical setting, it is assumed that the error term ε has the normal distribution with a mean 0 and a constant variance σ^2.

The first impression of the multiple regression may be a response plane. However, some regressors may be higher order terms of other regressors, or may even be functions of regressors as long as these functions do not contain unknown parameters. Thus, multiple regression model can be a response surface of versatile shapes. Readers may already realize the difference between a linear model and a nonlinear model.

Definition 3.1. A linear model is defined as a model that is linear in regression parameters, i.e., linear in β_i's.

The following are examples of linear regression models in which the response variable y is a linear function of regression parameters:

$$\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,\\
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon,\\
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \varepsilon,\\
y &= \beta_0 + \beta_1 x_1 + \beta_2 ln(x_1) + \beta_2 ln(x_2) + \varepsilon,\\
y &= \beta_0 + \beta_2 1_{(x_1>5)} + \beta_2 1_{(x_2>10)} + \beta_3 x_3 + \varepsilon.
\end{aligned}$$

In the last model $1_{(x_1>5)}$ is an indicator function taking value 1 if $x_1 > 5$ and 0 otherwise. Examples of non-linear regression model may be given by

$$\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2^\gamma + \varepsilon,\\
y &= \frac{1}{\lambda + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)} + \varepsilon,
\end{aligned}$$

where the response variable cannot be expressed as a linear function of regression parameters.

3.3 Quadratic Form of Random Variables

Definition 3.2. Let $y' = (y_1, y_2, \cdots, y_n)$ be n real variables and a_{ij} be $n \times n$ real numbers, where $i, j = 1, 2, \cdots, n$. A quadratic form of $y_1, y_2, \cdots, y_n$ is defined as

$$f(y_1, y_2, \cdots, y_n) = \sum_{i,j=1}^{n} a_{ij} y_i y_j.$$

This quadratic form can be written in the matrix form: $y' Ay$, where A is an $n \times n$ matrix $A = (a_{ij})_{n\times n}$. Quadratic form plays an important role in the discussions of linear regression model. In the classical setting the parameters of a linear regression model are estimated via minimizing the sum of squared residuals:

$$\begin{aligned}
&\mathbf{b} = (b_0, b_1, \cdots, b_k)\\
&= \mathbf{arg\,min}_{(\beta_0,\beta_1,\cdots,\beta_k)} \sum_{i=1}^{n} \Big[\, y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki})\Big]^2.
\end{aligned}$$

This squared residual is actually a quadratic form. Thus, it is important to discuss some general properties of this quadratic form that will be used in the subsequent discussions.

3.4 Idempotent Matrices

In this section we discuss properties of the idempotent matrix and its applications in the linear regression. First we define the idempotent matrix.

Definition 3.3. An $n \times n$ symmetric matrix A is idempotent if $A^2 = A$.

Let $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$ be a k-dimensional vector and A is a $k \times k$ matrix. $\alpha^{'} A\alpha$ is a quadratic form of $\alpha_1, \alpha_2, \cdots, \alpha_k$. When A is an idempotent matrix, the corresponding quadratic form has its particular properties. The quadratic form with idempotent matrices are used extensively in linear regression analysis. We now discuss the properties of idempotent matrix.

Theorem 3.4. *Let $A_{n\times n}$ be an idempotent matrix of rank p, then the eigenvalues of A are either 1 or 0.*

Proof. Let λ_i and v_i be the eigenvalue and the corresponding normalized eigenvector of the matrix A, respectively. We then have $Av_i = \lambda_i v_i$, and $v_i^{'} Av_i = \lambda_i v_i^{'} v_i = \lambda_i$. On the other hand, since $A^2 = A$, we can write

$$\lambda_i = v_i^{'} Av_i = v_i^{'} A^2 v_i = v_i^{'} A^{'} Av_i = (Av_i)^{'} Av_i = (\lambda_i v_i)^{'} (\lambda_i v_i) = \lambda_i^2.$$

Hence, we have $\lambda_i(\lambda_i - 1) = 0$, which yields either $\lambda_i = 1$ or $\lambda_i = 0$. This completes the proof. □

It is easy to know that p eigenvalues of A are 1 and $n - p$ eigenvalues of A are zero. Therefore, the rank of an idempotent matrix A is the sum of its non-zero eigenvalues.

Definition 3.4. Let $A = (a_{i,j})_{n\times n}$ be an $n\times n$ matrix, trace of A is defined as the sum of the orthogonal elements. i.e.,

$$tr(A) = a_{11} + a_{22} + \cdots + a_{nn}.$$

If A is a symmetric matrix then the sum of all squared elements of A can be expressed by $tr(A^2)$. i.e., $\sum_{i,j} a_{ij}^2 = tr(A^2)$. It is easy to verify that $tr(AB) = tr(BA)$ for any two $n \times n$ matrices A and B. The following theorem gives the relationship between the rank of matrix A and and trace of A when A is an idempotent matrix.

Theorem 3.5. *If A is an idempotent matrix then $tr(A) = rank(A) = p$.*

Proof. If the rank of an $n \times n$ idempotent matrix A is p then A has p eigenvalues of 1 and $n - p$ eigenvalues of 0. Thus, we can write $rank(A) =$

$\sum_{i=1}^{n} \lambda_i = p$. Since $A^2 = A$, the eigenvalues of the idempotent matrix A is either 1 or 0. From matrix theory there is an orthogonal matrix V such that

$$V^{'}AV = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}.$$

Therefore, we have

$$tr(V^{'}AV) = tr(VV^{'}A) = tr(A) = tr\begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix} = p = rank(A).$$

Here we use the simple fact: $tr(AB) = tr(BA)$ for any matrices $A_{n\times n}$ and $B_{n\times n}$. □

A quadratic form of a random vector $\boldsymbol{y}^{'} = (y_1, y_2, \cdots, y_n)$ can be written in a matrix form $\boldsymbol{y}^{'}A\boldsymbol{y}$, where A is an $n \times n$ matrix. It is of interest to find the expectation and variance of $\boldsymbol{y}^{'}A\boldsymbol{y}$. The following theorem gives the expected value of $\boldsymbol{y}^{'}A\boldsymbol{y}$ when the components of $\boldsymbol{y}$ are independent.

Theorem 3.6. *Let $\boldsymbol{y}^{'} = (y_1, y_2, \cdots, y_n)$ be an $n \times 1$ random vector with mean $\mu^{'} = (\mu_1, \mu_2, \cdots, \mu_n)$ and variance σ^2 for each component. Further, it is assumed that $y_1, y_2, \cdots, y_n$ are independent. Let A be an $n \times n$ matrix, $\boldsymbol{y}^{'}A\boldsymbol{y}$ is a quadratic form of random variables. The expectation of this quadratic form is given by*

$$E(\boldsymbol{y}^{'}A\boldsymbol{y}) = \sigma^2 tr(A) + \mu^{'}A\mu. \tag{3.6}$$

Proof. First we observe that

$$\boldsymbol{y}^{'}A\boldsymbol{y} = (\boldsymbol{y} - \mu)^{'}A(\boldsymbol{y} - \mu) + 2\mu^{'}A(\boldsymbol{y} - \mu) + \mu^{'}A\mu.$$

We can write

$$\begin{aligned} E(\boldsymbol{y}^{'}A\boldsymbol{y}) &= E[(\boldsymbol{y} - \mu)^{'}A(\boldsymbol{y} - \mu)] + 2E[\mu^{'}A(\boldsymbol{y} - \mu)] + \mu^{'}A\mu \\ &= E\Big[\sum_{i,j=1}^{n} a_{ij}(y_i - \mu_i)(y_j - \mu_j)\Big] + 2\mu^{'}AE(\boldsymbol{y} - \mu) + \mu^{'}A\mu \\ &= \sum_{i=1}^{n} a_{ii}E(y_i - \mu_i)^2 + \mu^{'}A\mu = \sigma^2 tr(A) + \mu^{'}A\mu. \end{aligned}$$

□

We now discuss the variance of the quadratic form $\boldsymbol{y}^{'}A\boldsymbol{y}$.

Theorem 3.7. *Let $\boldsymbol{y}$ be an $n \times 1$ random vector with mean $\mu^{'} = (\mu_1, \mu_2, \cdots, \mu_n)$ and variance σ^2 for each component. It is assumed that $y_1, y_2, \cdots, y_n$ are independent. Let A be an $n \times n$ symmetric matrix, $E(y_i - \mu_i)^4 = \mu_i^{(4)}$, $E(y_i - \mu_i)^3 = \mu_i^{(3)}$, and $a^{'} = (a_{11}, a_{22}, \cdots, a_{nn})$. The variance of the quadratic form $Y^{'}AY$ is given by*

$$\begin{aligned} Var(\boldsymbol{y}^{'}A\boldsymbol{y}) &= (\mu^{(4)} - 3\sigma^2)a'a + \sigma^4(2tr(A^2) + [tr(A)]^2) \\ &\quad +4\sigma^2\mu^{'}A^2\mu + 4\mu^{(3)}a^{'}A\mu. \end{aligned} \tag{3.7}$$

Proof. Let $Z = \boldsymbol{y} - \mu$, $A = (A_1, A_2, \cdots, A_n)$, and $b = (b_1, b_2, \cdots, b_n) = \mu^{'}(A_1, A_2, \cdots, A_n) = \mu^{'}A$ we can write

$$\begin{aligned} \boldsymbol{y}^{'}A\boldsymbol{y} &= (\boldsymbol{y}^{'} - \mu)A(\boldsymbol{y} - \mu) + 2\mu^{'}A(\boldsymbol{y} - \mu) + \mu^{'}A\mu \\ &= Z^{'}AZ + 2bZ + \mu^{'}A\mu. \end{aligned}$$

Thus

$$\text{Var}(\boldsymbol{y}^{'}A\boldsymbol{y}) = \text{Var}(Z^{'}AZ) + 4Var(bZ) + 4\text{Cov}(Z^{'}AZ,\ bZ).$$

We then calculate each term separately:

$$\begin{aligned} (Z^{'}AZ)^2 &= \sum_{ij} a_{ij}a_{lm}Z_iZ_jZ_lZ_m \\ E(Z^{'}AZ)^2 &= \sum_{i\ j\ l\ m} a_{ij}a_{lm}E(Z_iZ_jZ_lZ_m) \end{aligned}$$

Note that

$$E(Z_iZ_jZ_lZ_m) = \begin{cases} \mu^{(4)}, & \text{if } i = j = k = l; \\ \sigma^4, & \text{if } i = j,\ l = k \text{ or } i = l,\ j = k, \text{ or } i = k,\ j = l\ ; \\ 0, & \text{else.} \end{cases}$$

We have

$$\begin{aligned} E(Z^{'}AZ)^2 &= \sum_{i\ j\ l\ m} a_{ij}a_{lm}E(Z_iZ_jZ_lZ_m) \\ &= \mu^{(4)}\sum_{i=1}^{n} a_{ii}^2 + \sigma^4\Big(\sum_{i \neq k} a_{ii}a_{kk} + \sum_{i \neq j} a_{ij}^2 + \sum_{i \neq j} a_{ij}a_{ji}\Big) \end{aligned}$$

Since A is symmetric, $a_{ij} = a_{ji}$, we have

$$\begin{aligned}&\sum_{i\neq j} a_{ij}^2 + \sum_{i\neq j} a_{ij}a_{ji}\\ &= 2\sum_{i\neq j} a_{ij}^2 = 2\sum_{i,j} a_{ij}^2 - 2\sum_{i=j} a_{ij}^2\\ &= 2tr(A^2) - 2\sum_{i=1}^{n} a_{ii}^2\\ &= 2tr(A^2) - 2a^{'}a\end{aligned}$$

and

$$\begin{aligned}\sum_{i\neq k} a_{ii}a_{kk} &= \sum_{i,k} a_{ii}a_{kk} - \sum_{i=k} a_{ii}a_{kk}\\ = [tr(A)]^2 - \sum_{i=1}^{n} a_{ii}^2 &= [tr(A)]^2 - a^{'}a.\end{aligned}$$

So we can write

$$E(Z^{'}AZ)^2 = (\mu^{(4)} - 3\sigma^4)a^{'}a + \sigma^4(2tr(A^2) + [tr(A)]^2). \tag{3.8}$$

For Var(bZ) we have

$$\text{Var}(bZ) = b\text{Var}(Z)b^{'} = bb^{'}\sigma^2 = (\mu^{'}A)(\mu^{'}A)^{'}\sigma^2 = \mu^{'}A^2\mu\sigma^2. \tag{3.9}$$

To calculate Cov$(Z^{'}AZ,\ bZ)$, note that $EZ = 0$, we have

$$\begin{aligned}&\text{Cov}(Z^{'}AZ,\ bZ)\\ &= \text{Cov}\Big(\sum_{i,j} a_{ij}Z_iZ_j, \sum_{k} b_kZ_k\Big)\\ &= \sum_{i,j,k} a_{ij}b_k\text{Cov}(Z_iZ_j,\ Z_k)\\ &= \sum_{i,j,k} a_{ij}b_kE[(Z_iZ_j - E(Z_iZ_j))Z_k]\\ &= \sum_{i,j,k} a_{ij}b_k[E(Z_iZ_jZ_k) - E(Z_iZ_j)EZ_k]\\ &= \sum_{i,j,k} a_{ij}b_k[E(Z_iZ_jZ_k)] \quad \text{(since } EZ_k = 0).\end{aligned}$$

It is easy to know that

$$E(Z_iZ_jZ_k) = \begin{cases}\mu^{(3)}, & \text{if } i = j = k;\\ 0, & \text{else.}\end{cases}$$

Thus,

$$\begin{aligned}\text{Cov}(Z^{'}AZ,\ bZ) &= \sum_{i=1}^{n} a_{ii}b_i\mu^{(3)} \\ = \sum_{i=1}^{n} a_{ii}\mu^{'}A_i\mu^{(3)} &= \sum_{i=1}^{n} a_{ii}A_i^{'}\mu\ \mu^{(3)} = a^{'}A\mu\ \mu^{(3)}.\end{aligned} \tag{3.10}$$

Combining the results above completes the proof. □

3.5 Multivariate Normal Distribution

A random variable Y is said to follow the normal distribution $N(\mu, \sigma^2)$ if and only if the probability density function of Y is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ - \frac{(y-\mu)^2}{\sigma^2}\Big\} \text{ for } -\infty < y < \infty. \tag{3.11}$$

The cumulative distribution of Y is defined as

$$F(y) = P(Y \leq y) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{y} \exp\Big\{ - \frac{(y-\mu)^2}{\sigma^2}\Big\} dy. \tag{3.12}$$

The moment generating function for the normal random variable $Y \sim N(\mu,\ \sigma)$ is

$$M(t) = E(e^{tY}) = \exp(t\mu + \frac{1}{2}t^2\sigma^2). \tag{3.13}$$

The multivariate normal distribution is an extension of the univariate normal distribution. A random vector $\boldsymbol{y}^{'} = (y_1, y_2, \cdots, y_p)$ is said to follow the multivariate normal distribution if and only if its probability density function has the following form

$$\begin{aligned}& f(y_1, y_2, \cdots, y_p) \\ &= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\Big\{ -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{'}\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\Big\},\end{aligned} \tag{3.14}$$

where $\Sigma = (\sigma_{ij})_{p\times p}$ is the covariance matrix of $\boldsymbol{y}$ and the inverse matrix Σ^{-1} exists. $\boldsymbol{\mu}^{'} = (\mu_1, \mu_2, \cdots, \mu_p)$ is the mean vector of $\boldsymbol{y}$.

When Σ is a diagonal matrix $\Sigma = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2)$, or $\sigma_{ij} = 0$ for all $i \neq j$, then $y_1, y_2, \cdots, y_p$ are not correlated since it is easy to know that the density function of $\boldsymbol{y}$ can be written as a product of p univariate normal density function:

$$\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\Big\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{'}\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\Big\} = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\big\{-\frac{(y_i-\mu_i)^2}{\sigma_i^2}\big\}$$

Since density function of multivariate normal vector $\boldsymbol{y}$ is a product of density functions of $y_1, y_2, \cdots, y_p$, they are jointly independent. For multivariate normal variables, the uncorrelated normal random variables are jointly independent. We summarize this into the following theorem:

Theorem 3.8. *If random vector $\boldsymbol{y}^{'} = (y_1, y_2, \cdots, y_p)$ follows a multivariate normal distribution $N(\boldsymbol{\mu},\ \Sigma)$ and the covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$ is a diagonal matrix diag$(\sigma_{11}, \sigma_{22}, \cdots, \sigma_{pp})$, then $y_1, y_2, \cdots, y_p$ are jointly independent.*

We now introduce the central χ^2 distribution. Let $y_1, y_2, \cdots, y_p$ be p independent standard normal random variables, i.e., $E(y_i) = 0$ and $\mathrm{Var}(y_i) = 1$. The special quadratic form $Z = \sum_{i=1}^{p} y_i^2$ has the chi-square distribution with p degrees of freedom and non-centrality parameter $\lambda = 0$. In addition, the random variable Z has the density function

$$f(z) = \frac{1}{\Gamma(p/2)2^{p/2}} z^{(p-2)/2} e^{-z/2} \quad \text{for} \ \ 0 < z < \infty. \tag{3.15}$$

The moment generating function for Z is given by

$$M(t) = E(e^{tZ}) = (1-2t)^{-n/2} \quad \text{for} \ \ t < \frac{1}{2}. \tag{3.16}$$

Using this moment generating function it is easy to find $E(Z) = p$ and $\mathrm{Var}(Z) = 2p$. In addition, the following results are obtained through direct calculations:

$$E(Z^2) = p\,(p+2),$$

$$E(\sqrt{Z}\,) = \frac{\sqrt{2}\,\Gamma[(p+1)/2]}{\Gamma(p/2)},$$

$$E\Big(\frac{1}{Z}\Big) = \frac{1}{p-2},$$

$$E\Big(\frac{1}{Z^2}\Big) = \frac{1}{(n-2)(n-4)},$$

$$E\Big(\frac{1}{\sqrt{Z}}\Big) = \frac{\Gamma[(p-1/2)]}{\sqrt{2}\,\Gamma(p/2)}.$$

3.6 Quadratic Form of the Multivariate Normal Variables

The distribution of the quadratic form $\boldsymbol{y}^{'}A\boldsymbol{y}$ when $\boldsymbol{y}$ follows the multivariate normal distribution plays a significant role in the discussion of linear regression methods. We should further discuss some theorems about the distribution of the quadratic form based upon the mean and covariance matrix of a normal vector $\boldsymbol{y}$, as well as the matrix A.

Theorem 3.9. *Let $\boldsymbol{y}$ be an $n \times 1$ normal vector and $\boldsymbol{y} \sim N(0, I)$. Let A be an idempotent matrix of rank p. i.e., $A^2 = A$. The quadratic form $\boldsymbol{y}^{'}A\boldsymbol{y}$ has the chi-square distribution with p degrees of freedom.*

Proof. Since A is an idempotent matrix of rank p. The eigenvalues of A are 1's and 0's. Moreover, there is an orthogonal matrix V such that

$$VAV^{'} = \begin{pmatrix} I_p & 0 \\ 0 & 0 \end{pmatrix}.$$

Now, define a new vector $\boldsymbol{z} = V\boldsymbol{y}$ and $\boldsymbol{z}$ is a multivariate normal vector. $E(\boldsymbol{z}) = VE(\boldsymbol{y}) = 0$ and $\mathrm{Cov}(\boldsymbol{z}) = \mathrm{Cov}(V\boldsymbol{y}) = V\mathrm{Cov}(\boldsymbol{y})V^{'} = VI_pV^{'} = I_p$. Thus, $\boldsymbol{z} \sim N(0, I_p)$. Notice that V is an orthogonal matrix and

$$\boldsymbol{y}^{'}A\boldsymbol{y} = (V^{'}\boldsymbol{z})^{'}AV^{'}\boldsymbol{z} = \boldsymbol{z}^{'}VAV^{'}\boldsymbol{z} = \boldsymbol{z}^{'}I_p\boldsymbol{z} = \sum_{i=1}^{p} z_i^2.$$

By the definition of the chi-square random variable, $\sum_{i=1}^{p} z_i^2$ has the chi-square distribution with p degrees of freedom. □

The above theorem is for the quadratic form of a normal vector $\boldsymbol{y}$ when $E\boldsymbol{y} = 0$. This condition is not completely necessary. However, if this condition is removed, i.e., if $E(\boldsymbol{y}) = \mu \neq 0$ the quadratic form of $\boldsymbol{y}^{'}A\boldsymbol{y}$ still follows the chi-square distribution but with a non-centrality parameter $\lambda = \frac{1}{2}\boldsymbol{\mu}^{'}A\boldsymbol{\mu}$. We state the theorem and the proofs of the theorem should follow the same lines as the proofs of the theorem for the case of $\mu = 0$.

Theorem 3.10. *Let $\boldsymbol{y}$ be an $n \times 1$ normal vector and $y \sim N(\mu, I)$. Let A be an idempotent matrix of rank p. The quadratic form $\boldsymbol{y}^{'}A\boldsymbol{y}$ has the chi-square distribution with degrees of freedom p and the non-centrality parameter $\lambda = \frac{1}{2}\mu^{'}A\mu$.*

We now discuss more general situation where the normal vector $\boldsymbol{y}$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

Theorem 3.11. *Let $\boldsymbol{y}$ be a multivariate normal vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ. If $A\Sigma$ is an idempotent matrix of rank p, The quadratic form of $\boldsymbol{y}^{'}A\boldsymbol{y}$ follows a chi-square distribution with degrees of freedom p and non-centrality parameter $\lambda = \frac{1}{2}\mu^{'}A\mu$.*

Proof. First, for covariance matrix Σ there exists an orthogonal matrix Γ such that $\Sigma = \Gamma\Gamma^{'}$. Define $Z = \Gamma^{-1}(\boldsymbol{y} - \mu)$ and Z is a normal vector with $E(Z) = 0$ and

$$\begin{aligned}\text{Cov}(Z) &= \text{Cov}(\Gamma^{-1}(\boldsymbol{y} - \mu)) = \Gamma^{-1}\text{Cov}(\boldsymbol{y})\Gamma^{'-1} = \Gamma^{-1}\Sigma\Gamma^{'-1}\\ &= \Gamma^{-1}(\Gamma\Gamma^{'})\Gamma^{'-1} = I_p.\end{aligned}$$

i.e., $Z \sim N(0, I)$. Moreover, since $\boldsymbol{y} = \Gamma Z + \mu$ we have

$$\boldsymbol{y}^{'}A\boldsymbol{y} = [\Gamma Z + \mu)]^{'}A(\Gamma Z + \mu) = (Z^{'} + \Gamma^{'-1}\mu)^{'}(\Gamma^{'}A\Gamma)(Z + \Gamma^{'-1}\mu) = V^{'}BV,$$

where $V = Z^{'} + \Gamma^{'-1}\mu \sim N(\Gamma^{'-1}\mu,\ \ I_p)$ and $B = \Gamma^{'}A\Gamma$. We now need to show that B is an idempotent matrix. In fact,

$$B^2 = (\Gamma^{'}A\Gamma)(\Gamma^{'}A\Gamma) = \Gamma^{'}(A\Gamma\Gamma^{'}A)\Gamma$$

Since $A\Sigma$ is idempotent we can write

$$A\Sigma = A\Gamma\Gamma^{'} = A\Sigma A\Sigma = (A\Gamma\Gamma^{'}A)\Gamma\Gamma^{'} = (A\Gamma\Gamma^{'}A)\Sigma.$$

Note that Σ is non-singular we have

$$A = A\Gamma\Gamma^{'}A.$$

Thus,

$$B^2 = \Gamma^{'}(A\Gamma\Gamma^{'}A)\Gamma = \Gamma^{'}A\Gamma = B.$$

i.e., B is an idempotent matrix. This concludes that $V^{'}BV$ is a chi-square random variable with degrees of freedom p. To find the non-centrality parameter we have

$$\begin{aligned}\lambda &= \frac{1}{2}(\Gamma^{'-1}\mu)^{'}B(\Gamma^{'-1}\mu)\\ &= \frac{1}{2}\mu^{'}\Gamma^{'-1}(\Gamma^{'}A\Gamma)\Gamma^{'-1}\mu = \frac{1}{2}\mu^{'}A\mu.\end{aligned}$$

This completes the proof. □

3.7 Least Squares Estimates of the Multiple Regression Parameters

The multiple linear regression model is typically stated in the following form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i,$$

where y_i is the dependent variable, $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ are the regression coefficients, and ε_i's are the random errors assuming $E(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$ for $i = 1, 2, \cdots, n$. In the classical regression setting the error term is assumed to be normally distributed with a constant variance σ^2. The regression coefficients are estimated using the least squares principle. It should be noted that it is not necessary to assume that the regression error term follows the normal distribution in order to find the least squares estimation of the regression coefficients. It is rather easy to show that under the assumption of normality of the error term, the least squares estimation of the regression coefficients are exactly the same as the maximum likelihood estimations (MLE) of the regression coefficients.

The multiple linear model can also be expressed in the matrix format

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},$$

where

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_{k-1} \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdots \\ \varepsilon_n \end{pmatrix} \tag{3.17}$$

The matrix form of the multiple regression model allows us to discuss and present many properties of the regression model more conveniently and efficiently. As we will see later the simple linear regression is a special case of the multiple linear regression and can be expressed in a matrix format. The least squares estimation of $\boldsymbol{\beta}$ can be solved through the least squares principle:

$$\boldsymbol{b} = \mathbf{arg\,min}_{\boldsymbol{\beta}}[(\boldsymbol{y} - \boldsymbol{X\beta})^{'}(\boldsymbol{y} - \boldsymbol{X\beta})],$$

where $\boldsymbol{b}^{'} = (b_0, b_1, \cdots b_{k-1})^{'}$, a k-dimensional vector of the estimations of the regression coefficients.

Theorem 3.12. *The least squares estimation of $\boldsymbol{\beta}$ for the multiple linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ is $\boldsymbol{b} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}$, assuming $(\boldsymbol{X}^{'}\boldsymbol{X})$ is a non-singular matrix. Note that this is equivalent to assuming that the column vectors of $\boldsymbol{X}$ are independent.*

Proof. To obtain the least squares estimation of $\boldsymbol{\beta}$ we need to minimize the residual of sum squares by solving the following equation:

$$\frac{\partial}{\partial \boldsymbol{b}}[(\boldsymbol{y} - \boldsymbol{Xb})^{'}(\boldsymbol{y} - \boldsymbol{Xb})] = 0,$$

or equivalently,

$$\frac{\partial}{\partial \boldsymbol{b}}[(\boldsymbol{y}^{'}\boldsymbol{y} - 2\boldsymbol{y}^{'}\boldsymbol{Xb} + \boldsymbol{b}^{'}\boldsymbol{X}^{'}\boldsymbol{Xb})] = 0.$$

By taking partial derivative with respect to each component of $\boldsymbol{\beta}$ we obtain the following normal equation of the multiple linear regression model:

$$\boldsymbol{X}^{'}\boldsymbol{Xb} = \boldsymbol{X}^{'}\boldsymbol{y}.$$

Since $\boldsymbol{X}^{'}\boldsymbol{X}$ is non-singular it follows that $\boldsymbol{b} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}$. This completes the proof. □

We now discuss statistical properties of the least squares estimation of the regression coefficients. We first discuss the unbiasness of the least squares estimation $\boldsymbol{b}$.

Theorem 3.13. *The estimator $\boldsymbol{b} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}$ is an unbiased estimator of $\boldsymbol{\beta}$. In addition,*

$$Var(\boldsymbol{b}) = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\sigma^2. \tag{3.18}$$

Proof. We notice that

$$E\boldsymbol{b} = E((\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}) = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}E(\boldsymbol{y}) = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

This completes the proof of the unbiasness of $\boldsymbol{b}$. Now we further discuss how to calculate the variance of $\boldsymbol{b}$. The variance of the $\boldsymbol{b}$ can be computed directly:

$$\begin{aligned}
&\mathrm{Var}(\boldsymbol{b}) = \mathrm{Var}((\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y})\\
&= (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\mathrm{Var}(\boldsymbol{b})((\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})^{'}\\
&= (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\sigma^2 = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\sigma^2.
\end{aligned}$$

□

Another parameter in the classical linear regression is the variance σ^2, a quantity that is unobservable. Statistical inference on regression coefficients and regression model diagnosis highly depend on the estimation of error variance σ^2. In order to estimate σ^2, consider the residual sum of squares:

$$\boldsymbol{e}^t\boldsymbol{e} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^{'}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{y}^{'}[I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}]\boldsymbol{y} = \boldsymbol{y}^{'}P\boldsymbol{y}.$$

This is actually a distance measure between observed $\boldsymbol{y}$ and fitted regression value $\hat{\boldsymbol{y}}$. Note that it is easy to verify that $P = [I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}]$ is idempotent. i.e.,

$$P^2 = [I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}][I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}] = [I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}] = P.$$

Therefore, the eigenvalues of P are either 1 or 0. Note that the matrix $\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}$ is also idempotent. Thus, we have

$$\begin{aligned}
&rank(\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}) = tr(\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})\\
&= tr(\boldsymbol{X}^{'}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}) = tr(I_p) = p.
\end{aligned}$$

Since $tr(A-B)=tr(A)-tr(B)$ we have

$$\begin{aligned}&rank(I-\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})=tr(I-\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})\\ &=tr(I_n)-tr(\boldsymbol{X}^{'}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1})=n-p\end{aligned}$$

The residual of sum squares in the multiple linear regression is $\boldsymbol{e}'\boldsymbol{e}$ which can be written as a quadratic form of the response vector $\boldsymbol{y}$.

$$\boldsymbol{e}'\boldsymbol{e}=(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^{'}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})=\boldsymbol{y}^{'}(I-\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})\boldsymbol{y}.$$

Using the result of the mathematical expectation of the quadratic form we have

$$\begin{aligned}&E(\boldsymbol{e}'\boldsymbol{e})=E\Big[\boldsymbol{y}^{'}(I-\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})\boldsymbol{y}\Big]\\ &=(\boldsymbol{X}\beta)^{'}(I-\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})(\boldsymbol{X}\beta)+\sigma^2(n-p)\\ &=(\boldsymbol{X}\beta)^{'}(\boldsymbol{X}\beta-\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{X}\beta)+\sigma^2(n-p)=\sigma^2(n-p)\end{aligned}$$

We summarize the discussions above into the following theorem:

Theorem 3.14. *The unbiased estimator of the variance in the multiple linear regression is given by*

$$s^2=\frac{\boldsymbol{e}'\boldsymbol{e}}{n-p}=\frac{\boldsymbol{y}^{'}(I-\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})\boldsymbol{y}}{n-p}=\frac{1}{n-p}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2. \quad (3.19)$$

Let $P=\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}$. The vector $\boldsymbol{y}$ can be partitioned into two vectors $(I-P)\boldsymbol{y}=(I-\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})\boldsymbol{y}$ and $P\boldsymbol{y}=\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})\boldsymbol{y}$. Assuming the normality of regression error term $(I-P)\boldsymbol{y}$ is independent of $P\boldsymbol{y}$. To see this we simply calculate the covariance of $(I-P)\boldsymbol{y}$ and $P\boldsymbol{y}$:

$$\begin{aligned}&\mathrm{Cov}\Big((I-P)\boldsymbol{y},P\boldsymbol{y}\Big)\\ &=(I-P)\mathrm{Cov}(\boldsymbol{y})P=(I-P)P\sigma^2\\ &=(I-\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\sigma^2\\ &=[\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}-(\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}]\sigma^2\\ &=(\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}-\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})\sigma^2=0\end{aligned}$$

Since $(I-P)\boldsymbol{y}$ and $P\boldsymbol{y}$ are normal vectors, the zero covariance implies that they are independent of each other. Thus, the quadratic functions

$\boldsymbol{y}^{'}(I-P)\boldsymbol{y}$ and $\boldsymbol{y}^{'}P\boldsymbol{y}$ are independent as well. When P is idempotent, the quadratic function of a normal vector $\boldsymbol{y}^{'}P\boldsymbol{y}$ follows the chi-square distribution with degrees of freedom p, where $p = rank(P)$. This property can be used to construct the F test statistic that is commonly used in the hypothesis testing problem for multiple linear regression.

The above calculations can be simplified if we introduce the following theorem for the two linear transformations of a multivariate normal variable $\boldsymbol{y}$.

Theorem 3.15. *Let $\boldsymbol{y} \sim N(\mu,\ I)$ and A and B be two matrices. Two normal vectors $A\boldsymbol{y}$ and $B\boldsymbol{y}$ are independent if and only if $AB^{'} = 0$.*

Proof. Recall that the independence of two normal vectors is equivalent to zero covariance between them. We calculate the covariance of $A\boldsymbol{y}$ and $B\boldsymbol{y}$.

$$\text{Cov}(A\boldsymbol{y}, B\boldsymbol{y}) = A\text{Cov}(\boldsymbol{y})B^{'} = AB^{'}$$

Thus, the independence of two normal vectors $A\boldsymbol{y}$ and $B\boldsymbol{y}$ is equivalent to $AB^{'} = 0$. □

By using this theorem we can easily show that $(I-P)\boldsymbol{y}$ and $P\boldsymbol{y}$ are independent. In fact, because P is idempotent, therefore, $(I-P)P = P - P^2 = P - P = 0$. The result follows immediately.

3.8 Matrix Form of the Simple Linear Regression

The simple linear regression model is a special case of the multiple linear regression and can be expressed in the matrix format. In particular,

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}.$$

The formula for calculating $\boldsymbol{b}$ in matrix format can be applied to the simple linear regression.

$$\boldsymbol{X}^{'}\boldsymbol{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}.$$

It is not difficult to solve for $(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}$ analytically. In fact, the inverse matrix of $\boldsymbol{X}^{'}\boldsymbol{X}$ is given by

$$(\boldsymbol{X}^{'}\boldsymbol{X})^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

$$= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}.$$

The least squares estimation of the simple linear regression can be calculated based on its matrix form:

$$\boldsymbol{b} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y} == \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

$$= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ n\sum x_i y_i - \sum x_i \sum y_i \end{pmatrix}$$

$$= \begin{pmatrix} \dfrac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \\ \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{pmatrix}.$$

Thus, we have

$$\boldsymbol{b}_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \bar{y} - \boldsymbol{b}_1\bar{x}$$

and

$$\boldsymbol{b}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The results are exactly identical to the results derived in Chapter 2. The unbiasness of the $\boldsymbol{b}$ and the covariance of $\boldsymbol{b}$ can be shown for the simple linear regression using its matrix form as well. We left this to the readers.

3.9 Test for Full Model and Reduced Model

Before an appropriate linear regression model is chosen it is often unknown how many variables should be included in the regression model. A linear regression model with more variables may not always perform better than the regression model with less variables when both models are compared in terms of residual of sum squares. To compare two regression models in terms of the independent variables included in the models we need to test if the regression model with more independent variables performs statistically better than the regression model with less independent variables. To this end, we define the full regression model as:

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \tag{3.20}$$

and the reduced regression model as:

$$\boldsymbol{y} = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}. \tag{3.21}$$

A full linear regression model is the model with more independent variables and a reduced model is the model with a subset of the independent variables in the full model. In other words, the reduced regression model is the model nested in the full regression model. We would like to test the following hypothesis

$$H_0: \ \boldsymbol{\beta}_1 = 0 \ \text{ versus } \ H_1: \ \boldsymbol{\beta}_1 \neq 0.$$

Under the null hypothesis H_0, the error term of the regression model $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$. Denote $\boldsymbol{X} = (\boldsymbol{X}_1, \ \boldsymbol{X}_2)$, where $\boldsymbol{X}_1$ is an $n \times p_1$ matrix, $\boldsymbol{X}_2$ is an $n \times p_2$ matrix, and n is the total number of observations. A test statistic needs to be constructed in order to compare the full regression model with the reduced regression regression model. Consider the difference between the SSE of the full model and the SSE of the reduced model:

$$SSE_{reduced} = \boldsymbol{y}'(I - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2')\boldsymbol{y}$$

and

$$SSE_{full} = \boldsymbol{y}'(I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y},$$

$$SSE_{reduced} - SSE_{full} = \boldsymbol{y}'\Big(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\Big)\boldsymbol{y}.$$

The matrices $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$ are idempotent. In addition, it can be shown that the matrix $\Big(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' -$

$\boldsymbol{X}_2(\boldsymbol{X}_2^{'}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{'}\Big)$ is also idempotent and the rank of this matrix is p_1 which is the dimension of $\boldsymbol{\beta}_1$:

$$\begin{aligned}
&\text{Rank of } \Big(\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'} - \boldsymbol{X}_2(\boldsymbol{X}_2^{'}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{'}\Big) \\
&= tr\Big(\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'} - \boldsymbol{X}_2(\boldsymbol{X}_2^{'}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{'}\Big) \\
&= tr\Big(\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\Big) - tr\Big(\boldsymbol{X}_2(\boldsymbol{X}_2^{'}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{'}\Big) \\
&= tr(\boldsymbol{X}^{'}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}) - tr(\boldsymbol{X}_2^{'}\boldsymbol{X}_2(\boldsymbol{X}_2^{'}\boldsymbol{X}_2)^{-1}) \\
&= (p_1 + p_2) - p_2 = p_1
\end{aligned}$$

The distribution of the following quadratic form is the chi-square distribution with degrees of freedom p_1:

$$\boldsymbol{y}^{'}\Big(\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'} - \boldsymbol{X}_2(\boldsymbol{X}_2^{'}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{'}\Big)\boldsymbol{y} \sim \sigma^2\chi^2_{p_1}.$$

Note that the matrix $I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}$ is idempotent and its rank is $n - p_1$. Applying the theorem of the distribution of the quadratic form, it can be shown that total sum of residuals

$$s^2 = \boldsymbol{y}^{'}\Big(I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\Big)\boldsymbol{y} \sim \sigma^2\chi^2_{n-p},$$

where p is the total number of parameters. In addition, It can be shown that s^2 is independent of $SSE_{reduced} - SSE_{full}$. In fact, we only need to show

$$\Big[\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'} - \boldsymbol{X}_2(\boldsymbol{X}_2^{'}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{'}\Big]\Big[I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\Big] = 0.$$

It is easy to verify that

$$\Big[I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\Big]\Big[\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\Big] = 0.$$

It remains to show

$$\Big[I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\Big]\Big[\boldsymbol{X}_2(\boldsymbol{X}_2^{'}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{'}\Big] = 0.$$

It is straightforward that

$$\Big[I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\Big]\boldsymbol{X} = \Big[I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\Big](\boldsymbol{X}_1, \boldsymbol{X}_2) = 0.$$

Note that $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ we have

$$\Big[I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\Big]\boldsymbol{X}_2 = 0.$$

Therefore,

$$\Big[I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\Big]\boldsymbol{X}_2\Big[(\boldsymbol{X}_2^{'}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{'}\Big] = 0.$$

Thus, we can construct the following F test statistic:

$$F = \frac{\boldsymbol{y}^{'}\Big(\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'} - \boldsymbol{X}_2(\boldsymbol{X}_2^{'}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{'}\Big)\boldsymbol{y}/p_1}{\boldsymbol{y}^{'}\Big(I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\Big)\boldsymbol{y}/n-p} \sim F_{p_1, n-p}. \quad (3.22)$$

This test statistic can be used to test hypothesis $H_0 : \boldsymbol{\beta}_1 = 0$ versus $H_1 : \boldsymbol{\beta}_1 \neq 0$.

3.10 Test for General Linear Hypothesis

Consider the following multiple linear regression model

$$\boldsymbol{y} = \boldsymbol{\beta}\boldsymbol{X} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim N(0,\ \sigma^2 I_n)$. It may be of interest to test the linear function of model parameters. This can be formulated into the following general linear hypothesis testing problem:

$$H_0 : C\boldsymbol{\beta} = d \ \text{ versus } \ H_1 : C\boldsymbol{\beta} \neq d.$$

Here, C is a $r \times p$ matrix of rank r and $r \leq p$, p is the number of parameters in the regression model, or the dimension of $\boldsymbol{\beta}$. Suppose that $\boldsymbol{b}$ is the least squares estimation of $\boldsymbol{\beta}$ then we have

$$\boldsymbol{b} \sim N(\boldsymbol{\beta},\ \sigma^2(\boldsymbol{X}^{'}\boldsymbol{X})^{-1})$$

and

$$C\boldsymbol{b} \sim N(C\boldsymbol{\beta},\ \sigma^2 C(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}C^{'}).$$

Under the null hypothesis $H_0 : C\boldsymbol{\beta} = d$, we have

$$[C\boldsymbol{b} - d\,]^{'}[C(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}C^{'}]^{-1}[C\boldsymbol{b} - d\,] \sim \sigma^2\chi^2_r,$$

therefore, the statistic that can be used for testing $H_0 : C\boldsymbol{\beta} = d$ versus $H_1 : C\boldsymbol{\beta} \neq d$ is the F test statistic in the following form:

$$F = \frac{(C\boldsymbol{b} - d)^{'}[C(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}C^{'}]^{-1}(C\boldsymbol{b} - d)}{rs^2} \sim F_{r,n-p}. \tag{3.23}$$

3.11 The Least Squares Estimates of Multiple Regression Parameters Under Linear Restrictions

Sometimes, we may have more knowledge about regression parameters, or we would like to see the effect of one or more independent variables in a regression model when the restrictions are imposed on other independent variables. This way, the parameters in such a regression model may be useful for answering a particular scientific problem of interest. Although restrictions on regression model parameters could be non-linear we only deal with the estimation of parameters under general linear restrictions. Consider a linear regression model

$$\boldsymbol{y} = \boldsymbol{\beta}\boldsymbol{X} + \boldsymbol{\varepsilon}.$$

Suppose that it is of interest to test the general linear hypothesis: $H_0 : C\boldsymbol{\beta} = d$ versus $H_1 : C\boldsymbol{\beta} \neq d$, where d is a known constant vector. We would like to explore the relationship of SSEs between the full model and the reduced model. Here, the full model is referred to as the regression model without restrictions on the parameters and the reduced model is the model with the linear restrictions on parameters. We would like to find the least squares estimation of $\boldsymbol{\beta}$ under the general linear restriction $C\boldsymbol{\beta} = d$. Here C is a $r \times p$ matrix of rank r and $r \leq p$. With a simple linear transformation the general linear restriction $C\boldsymbol{\beta} = d$ can be rewritten as $C\boldsymbol{\beta}^* = 0$. So, without loss of generality, we consider homogeneous linear restriction: $C\boldsymbol{\beta} = 0$. This will simplify the derivations. The estimator we are seeking for will minimize the least squares $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{'}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ under the linear restriction $C\boldsymbol{\beta} = 0$. This minimization problem under the linear restriction can be solved by using the method of the Lagrange multiplier. To this end, we construct the objective function $Q(\boldsymbol{\beta}, \lambda)$ with Lagrange multiplier λ:

$$
\begin{aligned}
Q(\boldsymbol{\beta}, \lambda) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{'}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\lambda C\boldsymbol{\beta} \\
&= \boldsymbol{y}^{'}\boldsymbol{y} + \boldsymbol{\beta}^{'}\boldsymbol{X}^{'}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^{'}\boldsymbol{X}^{'}\boldsymbol{y} - \boldsymbol{y}^{'}\boldsymbol{X}\boldsymbol{\beta} + 2\lambda C\boldsymbol{\beta}
\end{aligned}
$$

To minimize the objective function $Q(\boldsymbol{\beta}, \lambda)$, we take the partial derivatives with respect to each component of $\boldsymbol{\beta}$ and with respect to λ which yields the following normal equations:

$$
\begin{cases}
\boldsymbol{X}^{'}\boldsymbol{X}\boldsymbol{\beta} + C\lambda = \boldsymbol{X}^{'}\boldsymbol{y} \\
C\boldsymbol{\beta} = 0
\end{cases}
$$

The solutions of the above normal equation are least squares estimators of the regression model parameters under the linear restriction $C\boldsymbol{\beta} = 0$. The normal equation can be written in the form of blocked matrix:

$$
\begin{pmatrix} \boldsymbol{X}^{'}\boldsymbol{X} & C^{'} \\ C & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \lambda \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^{'}\boldsymbol{y} \\ 0 \end{pmatrix} \tag{3.24}
$$

The normal equation can be easily solved if one can find the inverse matrix on the left of the above normal equation. Formula of inverse blocked matrix can be used to solve the solution of the system. To simplify the notations we denote $\boldsymbol{X}^{'}\boldsymbol{X} = A$, and the inverse matrix in blocked form is given by

$$
\begin{pmatrix} \boldsymbol{X}^{'}\boldsymbol{X} & C^{'} \\ C & 0 \end{pmatrix}^{-1} = \begin{pmatrix} A & C^{'} \\ C & 0 \end{pmatrix}^{-1}
$$

$$
= \begin{pmatrix} A^{-1} - A^{-1}C^{'}(CA^{-1}C^{'})^{-1}CA^{-1} & A^{-1}C(CA^{-1}C^{'})^{-1} \\ & \\ (CA^{-1}C^{'})^{-1}C^{'}A^{-1} & -(CA^{-1}C^{'})^{-1} \end{pmatrix}
$$

By multiplying the blocked inverse matrix on the both sides of the above normal equation the least squares estimator of $\boldsymbol{\beta}$ under the linear restriction is given by

$$
\boldsymbol{b}^{*} = (A^{-1} - A^{-1}C^{'}(CA^{-1}C^{'})^{-1}CA^{-1})\boldsymbol{X}^{'}\boldsymbol{y}. \tag{3.25}
$$

For the full model (the model without restriction)

$$
SSE_{full} = \boldsymbol{y}^{'}(I - \boldsymbol{X}A^{-1}\boldsymbol{X}^{'})\boldsymbol{y}.
$$

For the reduced model (the model with a linear restriction):

$$SSE_{red} = \boldsymbol{y}^{'}(I - \boldsymbol{X}A^{-1}\boldsymbol{X}^{'} + \boldsymbol{X}A^{-1}C^{'}(CA^{-1}C^{'})^{-1}CA^{-1}\boldsymbol{X}^{'})\boldsymbol{y}$$

Note that $\boldsymbol{b} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}$ and we have

$$\begin{aligned} & SSE_{red} - SSE_{full} = \boldsymbol{y}^{'}(\boldsymbol{X}A^{-1}C(CA^{-1}C^{'})^{-1}CA^{-1}\boldsymbol{X}^{'})\boldsymbol{y} \\ = \ & \boldsymbol{y}^{'}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}C^{'}(C(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}C^{'})^{-1}C(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y} \\ = \ & (C\boldsymbol{b})^{'}(C(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}C^{'})^{-1}C\boldsymbol{b}. \end{aligned}$$

Under the normality assumption the above expression is a quadratic form of the normal variables. It can be shown that it has the chi-square distribution with degrees of freedom as the rank of the matrix $C(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}C^{'}$, which is r, the number of parameters in the model. Thus, we can write

$$(C\boldsymbol{b})^{'}[C(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}C^{'}]^{-1}(C\boldsymbol{b}) \sim \sigma^2\chi^2_r. \tag{3.26}$$

It can be shown that the s^2 is independent of the above χ^2 variable. Finally, we can construct the F test statistic:

$$F = \frac{(C\boldsymbol{b})^{'}[C(\boldsymbol{X}^{'}\boldsymbol{X})^{-1})C^{'}]^{-1}(C\boldsymbol{b})}{rs^2} \sim F_{r,\, n-p}, \tag{3.27}$$

which can be used to test the general linear hypothesis $H_0 : C\boldsymbol{\beta} = 0$ versus $H_1 : C\boldsymbol{\beta} \neq 0$.

3.12 Confidence Intervals of Mean and Prediction in Multiple Regression

We now discuss the confidence intervals on regression mean and regression prediction for multiple linear regression. For a given data point $\boldsymbol{x}_0^{'}$ the fitted value is $\hat{y}|\boldsymbol{x}_0 = \boldsymbol{x}_0^{'}\boldsymbol{b}$ and $Var(\hat{y}|\boldsymbol{x}_0) = \boldsymbol{x}_0^{'}\text{Cov}(\boldsymbol{b})\boldsymbol{x}_0 = \boldsymbol{x}_0^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{x}_0\sigma^2$. Note that under the normality assumption on the model error term $E(\hat{y}|\boldsymbol{x}_0) = E(x_0\boldsymbol{b}) = x_0^{'}\boldsymbol{\beta}$ and

$$\frac{(\hat{y}|\boldsymbol{x}_0) - E(\hat{y}|\boldsymbol{x}_0)}{s\sqrt{\boldsymbol{x}_0^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim t_{n-p}$$

where n is the total number of observations and p is the number of the parameters in the regression model. Thus, the $(1-\alpha)100\%$ confidence interval for $E(\hat{y}|\boldsymbol{x}_0)$ is given by

$$(\hat{y}|\boldsymbol{x}_0) \pm t_{\alpha/2,n-p} s \sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0} \qquad (3.28)$$

Using the arguments similar to that in Chapter 2 the confidence interval on regression prediction in multiple linear regression is given by:

$$(\hat{y}|\boldsymbol{x}_0) \pm t_{\alpha/2,n-p} s \sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0} \qquad (3.29)$$

3.13 Simultaneous Test for Regression Parameters

Instead of testing for regression parameters individually, we can simultaneously test for the model parameters. We describe this simultaneous hypothesis test for multiple regression parameters using the vector notation:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0, \text{ versus } H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0,$$

where $\boldsymbol{\beta}' = (\beta_0, \beta_1, \cdots, \beta_{p-1})$, a p-dimensional vector of regression parameters, and $\boldsymbol{\beta}_0' = (\beta_{00}, \beta_{10}, \cdots, \beta_{p-1,0})$, a p-dimensional constant vector. The above simultaneous hypothesis testing problem can be tested using the following F test statistic which has the F distribution with degrees of freedom p and $n-p$ under H_0:

$$F = \frac{(\boldsymbol{b} - \boldsymbol{\beta}_0)'(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{b} - \boldsymbol{\beta}_0)}{ps^2} \sim F_{p,n-p}.$$

Here n is the total number of observations and p is the total number of regression parameters. To test simultaneously the regression parameters, for a given test level α, if the observed $\boldsymbol{b}$ satisfies the following inequality for a chosen cut-off $F_{\alpha,p,n-p}$,

$$Pr\Big(\frac{(\boldsymbol{b} - \boldsymbol{\beta}_0)'(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{b} - \boldsymbol{\beta}_0)}{ps^2} \leq F_{\alpha,p,n-p}\Big) \geq 1 - \alpha,$$

then H_0 cannot be rejected. Otherwise, we accept H_1. The F test statistic can be used to construct the simultaneous confidence region for regression parameters $\boldsymbol{\beta}$:

$$\Big\{\boldsymbol{\beta} : \frac{(\boldsymbol{b} - \boldsymbol{\beta})'(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{b} - \boldsymbol{\beta})}{ps^2} \leq F_{\alpha,p,n-p}\Big\}. \qquad (3.30)$$

Note that this simultaneous confidence region of the regression parameters is an ellipsoid in $\mathbb{R}^p$.

3.14 Bonferroni Confidence Region for Regression Parameters

Instead of constructing an ellipsoid confidence region by a quadratic form of the regression coefficients we can set a higher confidence level for each parameter so that the joint confidence region for all regression coefficients has a confidence level $(1-\alpha)100\%$. This can be done using the Bonferroni approach. Suppose that we have p regression coefficients and would like to construct a $(1-\alpha)100\%$ joint confidence region for p regression parameters, instead of using $\alpha/2$ for each regression parameter we now use a higher level $\alpha/2p$ to construct the Bonferroni confidence interval for all regression parameters β_i, $i = 1, 2, \cdots, p$. i.e., we choose a cut-off $t_{\alpha/2p,\, n-p}$ and construct the following confidence interval for each regression parameter:

$$b_i \pm t_{\alpha/2p,\, n-p}(\text{standard error of } b_i).$$

Note that $\text{Cov}(\boldsymbol{b}) = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\sigma^2$. The standard error of b_i can be estimated by the squared root of the diagonal elements in the matrix $(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}s^2$. This confidence region is the p-dimensional rectangular in $\mathbb{R}^p$ and has a joint confidence level of not less than $1-\alpha$. Confidence region based on the Bonferroni approach is conservative but the calculation is simpler.

The Bonferroni method can also be used to construct the confidence bounds on regression mean. Suppose we have r data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_r$, and want to construct the Bonferroni simultaneous confidence intervals on regression means at points $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_r$. The following formula gives the simultaneous confidence intervals for regression means at the observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_r$:

$$\hat{y}(x_j) \pm t_{\alpha/2r,\, n-p}\, s\sqrt{x_j^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_j} \tag{3.31}$$

The SAS code for calculating simultaneous confidence intervals on regression means and regression predictions are similar to those for the simple linear regression which was presented in the previous chapter for the simple linear regression. The only difference is to set a higher confidence level $(1-\alpha/2r)100\%$ for the simultaneous confidence intervals on regression means at $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_r$.

3.15 Interaction and Confounding

We have seen that linear regression is flexible enough to incorporate certain nonlinearity in independent variables via polynomial or other transformed terms of the independent variables. It is quite useful in many applications that the independent variables in the linear regression are categorical. In this section, we shall continue to demonstrate its great flexibility in handling and exploring interactions. We will also show how linear regression is used to evaluate the confounding effects among predictors.

A confounding variable (or factor) is an extraneous variable in a regression model that correlates (positively or negatively) with both the dependent variable and the independent variable. A confounding variable is associated with both the probable cause and the outcome. In clinical study, the common ways of experiment control of the confounding factor are case-control studies, cohort studies, and stratified analysis. One major problem is that confounding variables are not always known or measurable. An interaction in a regression model often refers to as the effect of two or more independent variables in the regression model is not simply additive. Such a term reflects that the effect of one independent variable depends on the values of one or more other independent variables in the regression model.

The concepts of both interaction and confounding are more methodological than analytic in statistical applications. A regression analysis is generally conducted for two goals: to predict the response Y and to quantify the relationship between Y and one or more predictors. These two goals are closely related to each other; yet one is more emphasized than the other depending on application contexts. For example, in spam detection, prediction accuracy is emphasized as determining whether or not an incoming email is a spam is of primary interest. In clinical trials, on the other hand, the experimenters are keenly interested to know if an investigational medicine is more effective than the *control* or *exposure*, for which the standard treatment or a placebo is commonly used, in treating some disease. The assessment of treatment effect is often desired in analysis of many clinical trials. Both interaction and confounding are more pertaining to the second objective.

Consider a regression analysis involving assessment of the association between the response and one (or more) predictor, which may be affected by other extraneous predictors. The predictor(s) of major interest can be either categorical or continuous. When it is categorical, it is often referred to as *treatment* in experimental designs. The difference it makes on the responses is cited as the *treatment effect.* The extraneous predictors that potentially influence the treatment effect are termed as *covariates* or *control variables.* Interaction and confounding can be viewed as different manners in which the covariates influence the treatment effect.

3.15.1 *Interaction*

By definition, *interaction* is referred to as the situation where the association of major concern or the treatment effect varies with the levels or values of the covariates. Consider, for example, the treatment-by-center interaction in a multi-center clinical trial, a common issue involved in a clinical trial that is conducted in different medical centers. If the treatment effect remains the same among different medical centers, then we say that no interaction exists between treatment and center; if the treatment is found effective, nevertheless, more or less across different medical centers, then we say interaction exists between treatment and center and interaction involved is referred to as *quantitative interaction*; if the new treatment is found effective than the control in some medical centers but harmful than the control in some other centers, then the interaction is referred to as *qualitative interaction.* There is a directional change in treatment effect across centers in the case of qualitative interaction while the treatment effect only differs in amount, not in direction of the comparison, with quantitative interactions. Quantitative interactions are quite common. But if qualitative interaction exists, it causes much more concerns. It is thus imperative in clinical trials to detect and, if exists, fully explore and test for qualitative interaction. In the following discussion, we shall treat these two types of interaction by the same token, while referring interested readers to Gail and Simon (1985) and Yan and Su (2005) for more discussion on their important differences.

In linear regression, interaction is commonly formulated by cross-product terms. Consider the regression setting of response Y and two continuous regressors X_1 and X_2. The interaction model can be stated as, ignoring the subscript i for observations,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon. \qquad (3.32)$$

Recall that in the additive or main effect model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \qquad (3.33)$$

the association between Y and X_1 is mainly carried by its slope β_1, which corresponds to the amount of change in the mean response $E(Y)$ with one unit increase in X_1, holding X_2 fixed. Here the slope β_1, which does not depend on X_2, remains unchanged with different values of X_2 and hence can be interpreted as the *main effect* of X_1. Similar interpretation holds for the slope β_1 of X_2.

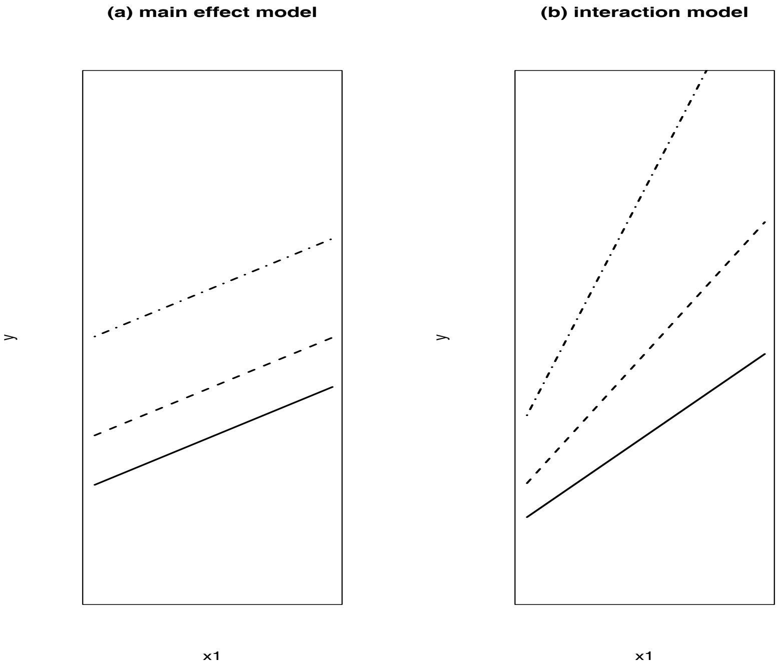

Fig. 3.1 Response Curves of Y Versus X_1 at Different Values of X_2 in Models (3.33) and (3.32).

In model (3.32), we can extract the 'slopes' for X_1 and X_2 by rewriting

$$\begin{aligned} E(y) &= (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) \cdot x_1 \\ &= (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1) \cdot x_2 \end{aligned}$$

The slope for X_1 now becomes $(\beta_1 + \beta_3 x_2)$, which depends on what value X_2 is fixed at. For this reason, X_2 is said to be an effect-modifier of X_1.

It is instructive to plot the response curves for Y versus X_1 at different values of X_2 for models (3.33) and (3.32), as shown in Fig. 3.1. We can see that the response curves in the main effect model are parallel lines with the same slope and different intercepts while in the interaction model the lines are no longer parallel. This explains why no interaction is often viewed as synonymous to parallelism, a principle in interaction detection that is applicable to various settings such as two-way analysis of variance (ANOVA) and comparing two or more response curves. Analogously, the slope for X_2 is $(\beta_2 + \beta_3 x_1)$, which depends on what value X_1 is fixed at.

Interaction among predictors can be generally formulated as cross product terms. For instance, an interaction model for Y versus X_1, X_2, and X_3 can be written as

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3 + \varepsilon \\ &= (\beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_6 x_2 x_3) + (\beta_1 + \beta_4 x_2 + \beta_5 x_3 + \beta_7 x_2 x_3)\, x_1 + \varepsilon. \end{aligned}$$

The products involving two terms $x_i x_j$, $i \neq j$, are referred to as first-order interactions; the three-term cross-products such as $x_i x_j x_k$, $i \neq j \neq k$, are called second-order interactions; and so on for higher-order interactions in general. The higher order of the interaction, the more difficult it would be in model interpretation. As seen in the above model, the slope for X_1 is $(\beta_1 + \beta_4 x_2 + \beta_5 x_3 + \beta_7 x_2 x_3)$, which depends on both x_2 and x_3 values in a complicated manner. To retain meaningful and simple model interpretation, it is often advised to consider interactions only up to the second order. In reality, interaction can be of high order with a complicated form other than cross products, which renders interaction detection a dunting task sometimes.

3.15.2 *Confounding*

Confounding is generally related to the broad topic of variable controlling or adjustment. Variable controlling and adjustment, which plays an important role to help prevent bias and reduce variation in treatment effect assessment, can be incorporated into a study at two stages. The first stage is in the design of the study. Consider, for instance, a study where the objective is to compare the prices of soft drinks of different brands, say, (A, B, and C). In a completely randomized design, one randomly goes to a number of grocery stores, pick up a drink of Brand A from each store, and record its price; then another set of grocery stores are randomly selected for Brand B; and so on for Brand C. Data collected in this manner result in several independent random samples, one for each treatment or brand, and the analysis can be carried out using the one-way ANOVA technique. The potential problem with this design, however, is that the treatment effect, as measured by the differences in price among the three groups, would be contaminated due to heterogeneity in other factors. Imagine what would happen if it turns out that price data collected for Brand A are taken from stores, mostly located in Minnesota in winter times while data collected for Brand B are taken during summer times from stores mostly located in Florida. In this case, we will be unable to obtain a genuine evaluation of the price difference due to brands. A better approach in this study is to employ a randomized block design with grocery stores being blocks, which can be described as follows. One randomly selects a number of grocery stores first; at each store, pick up a Brand A drink, a Brand B drink, and a Brand C drink and record their prices. In this way, we are able to control for many other geographical and longitudinal factors. By controlling, it means to make sure that they have the same or similar values. In general, if we know which factors are potentially important, then we can make control for them beforehand by using block or stratified designs.

However, very often we are pretty much blind about which factors are important. Sometimes, even if we have a good idea about them according to previous studies or literatures, we nevertheless do not have the authority or convenience to perform the control beforehand in the design stage. This is the case in many *observational studies.* Or perhaps there are too many of them; it is impossible to control for all. In this case, the adjustment can still be made in a *post hoc* manner at the data analysis stage. This is exactly what the analysis of covariance (ANCOVA) is aimed for. The approach is to fit models by including the important covariates.

The conception of confounding is casted into the *post hoc* variable adjustment at the data analysis stage. In general, *confounding* occurs if interpretations of the treatment effect are statistically different when some covariates are excluded or included in the model. It is usually assessed through a comparison between a crude estimator of the treatment effect by ignoring the extraneous variables and an estimate after adjusting for the covariates. Consider a setting (Y vs. $Z, X_1, \ldots, X_p$), where variable Z denotes the treatment variable of major interest and $X_1, \ldots, X_p$ denote the associated covariates. The comparison can be carried out in terms of the following two models:

$$y = \beta_0 + \beta_1 z + \varepsilon \tag{3.34}$$

and

$$y = \beta_0 + \beta_1 z + \alpha_1 x_1 + \cdots + \alpha_p x_p + \varepsilon. \tag{3.35}$$

Let $\hat{\beta}_1^{(c)}$ denote the least squares estimator of β_1 in model (3.34), which gives a rudimentary assessment of the treatment effect. Let $\hat{\beta}_1^{(a)}$ denote the least squares estimator of β_1 in model (3.35), which evaluates the treatment effect after adjusting or controlling for covariates $(X_1, \ldots, X_p)$. We say confounding is present if these two estimates, combined with their standard errors, are statistically different from each other. In this case, $(X_1, \ldots, X_p)$ are called confounders (or confounding factors) of Z.

In the traditional assessment of confounding effects, a statistical test is not required, perhaps because the analytical properties of $(\hat{\beta}_1^{(a)} - \hat{\beta}_1^{(c)})$ are not easy to comprehend unless resampling techniques such as bootstrap is used. It is mainly up to field experts to decide on existence of confounders and hence can be subjective. Another important point about confounding is that its assessment would become irrelevant if the treatment is strongly interacted with covariates. Interaction should be assessed before looking for confounders as it no longer makes sense to purse the main or separate effect of the treatment when it really depends on the levels or values of the covariates.

3.16 Regression with Dummy Variables

In regression analysis, a dummy variable is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift outcome. The reason we say "dummy" because it is not a variable that carries value of actual magnitude. For example, in a clinical trial, it is often useful to define a dummy variable D and $D = 1$ represents treatment group and $D = 0$ indicates placebo group; or we can introduce dummy variable S and define $S = 1$ for male group and 0 for female group. The mean value of a dummy variable is the proportion of the cases in the category coded 1. The variance of a dummy variable is $\sum D_i^2/n - (\sum D_i/n)^2 = p - p^2 = p(1-p)$, where p is the proportion of the cases in the category coded 1. Example of a regression model with dummy variable gender is:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \varepsilon_i \tag{3.36}$$

where Y_i is the annual salary of a lawyer, $D_i = 1$ if the lawyer is male and $D_i = 0$ if the lawyer is female, and X_i is years of experience. This model assumes that there is a mean salary shift between male lawyers and female lawyers. Thus, the mean salary is $E(Y_i|D_i = 1) = \beta_0 + \beta_2 + \beta_1 X_i$ for a male lawyer and is $E(Y_i|D_i = 0) = \beta_0 + \beta_1 X_i$ for a female lawyer. A test of the hypothesis $H_0 : \beta_2 = 0$ is a test of the hypothesis that the wage is the same for male lawyers and female lawyers when they have the same years of experience.

Several dummy variables can be used together to deal with more complex situation where more than two categories are needed for regression analysis. For example, variable race usually refers to the concept of categorizing humans into populations on the basis of various sets of characteristics. A variable race can have more than two categories. Suppose that we wanted to include a race variable with three categories White/Asian/Black in a regression model. We need to create a whole new set of dummy variables as follows

$$\begin{cases} D_{1i} = 1, \text{ if the person is white} \\ D_{1i} = 0, \text{ otherwise} \\ D_{2i} = 1, \text{ if the person is asian} \\ D_{2i} = 0, \text{ otherwise} \end{cases}$$

Here, the 'black' person category is treated as the base category and there is no need to create a dummy variable for this base category. All salary comparisons between two races in the regression model will be relative to the base category. In general, if there are m categories that need to be considered in a regression model it is needed to create $m - 1$ dummy variables, since the inclusion of all categories will result in perfect collinearity. Suppose that we would like to model the relation between the salary of a lawyer in terms of years of experience (X_i) and his/her race determined jointly by two dummy variables D_{1i}, D_{2i}, we can use the following regression model with the two dummy variables :

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \varepsilon_i, \tag{3.37}$$

where Y_i is the salary of the lawyer, X_i is years of his/her working experience, and D_{1i} and D_{2i} are dummy variables that determine the race of the lawyer. For example, based on the above regression model, the expected salary for a black lawyer with X_i years of working experience is

$$E(Y_i|D_{1i} = 0, D_{2i} = 0) = \beta_0 + \beta_1 X_i.$$

The expected salary for a white lawyer with X_i years of working experience is

$$E(Y_i|D_{1i} = 1, D_{2i} = 0) = \beta_0 + \beta_1 X_i + \beta_2.$$

The expected salary for an asian lawyer with X_i years of working experience is

$$E(Y_i|D_{1i} = 0, D_{2i} = 1) = \beta_0 + \beta_1 X_i + \beta_3.$$

In each case the coefficient of the dummy variable in the regression model represents the difference with the base race (the black lawyer's salary). Thus, the interpretation of β_2 is that a white lawyer earns β_2 more than a black lawyer, and the interpretation of β_3 is that an asian lawyer earns β_3 more than a black lawyer. The hypothesis test $H_0 : \beta_2 = 0$ is to test whether the wage is identical for a white lawyer and a black lawyer with same years of experience. And the hypothesis test $H_0 : \beta_3 = 0$ is to test that whether the wage is identical for an asian lawyer and a black lawyer with same years of experience.

Furthermore, if we would like to consider race effect and gender effect together, the following model with multiple dummy variables can be used:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 D_i + \varepsilon_i \tag{3.38}$$

According to model (3.38), for example, the expected salary for a female black lawyer with X_i years of experience is

$$E(Y_i|D_{1i} = 0, D_{2i} = 0, D_i = 0) = \beta_0 + \beta_1 X_i.$$

For a black male lawyer with X_i years of experience, the expected salary is

$$E(Y_i|D_{1i} = 0, D_{2i} = 0, D_i = 1) = \beta_0 + \beta_1 X_i + \beta_4.$$

For a white male lawyer with X_i years of experience, the expected salary is

$$E(Y_i|D_{1i} = 1, D_{2i} = 0, D_i = 1) = \beta_0 + \beta_1 X_i + \beta_2 + \beta_4.$$

The hypothesis test $H_0 : \beta_4 = 0$ is to test if the wage is the same for male lawyer and female lawyer with the same years of experience. The hypothesis test $H_0 : \beta_2 = \beta_4 = 0$ is to test if the wage is the same for male white lawyer and black lawyer with the same years of experience and if the gender has no impact on the salary.

If we have k categories then $k - 1$ dummy variables are needed. This is because in the classical regression it is required that none of exploratory variable should be a linear combination of remaining exploratory model variables to avoid collinearity. For example, we can use the dummy variable D and code $D = 1$ for male, if we also use another dummy variable $S = 0$ to indicate female, then there is a linear relation between D and S: $D = 1 - S$. Therefore, information become redundant. Thus, one dummy variable should be sufficient to represent information on gender. In general, $k - 1$ dummy variables are sufficient to represent k categories. Note that if D_i's, $i = 1, 2, \cdots, k - 1$, are $k - 1$ dummy variables then $D_i = 1$ represents a category out of the total k categories, and all $D_i = 0$ represents the base category out of the total k categories. Thus, $k - 1$ dummy variables are sufficient to represent k distinct categories.

If a regression model involves a nominal variable and the nominal variable has more than two levels, it is needed to create multiple dummy variables to replace the original nominal variable. For example, imagine that you wanted to predict depression level of a student according to status of freshman, sophomore, junior, or senior. Obviously, it has more than two levels. What you need to do is to recode "year in school" into a set of dummy variables, each of which has two levels. The first step in this process is to decide the number of dummy variables. This is simply $k - 1$, where k is the number of levels of the original nominal variable. In this instance, 3 dummy variables are needed to represent 4 categories of student status.

In order to create these variables, we are going to take 3 levels of "year in school", and create a variable corresponding to each level, which will have the value of yes or no (i.e., 1 or 0). In this example, we create three variables sophomore, junior, and senior. Each instance of "year in school" would then be recoded into a value for sophomore, junior, and senior. If a person is a junior, then variables sophomore and senior would be equal to 0, and variable junior would be equal to 1. A student with all variables sophomore, junior, and senior being all 0 is a freshman.

The decision as to which level is not coded is often arbitrary. The level which is not coded is the category to which all other categories will be compared. As such, often the biggest group will be the not-coded category. In a clinical trial often the placebo group or control group can be chosen as the not-coded group. In our example, freshman was not coded so that we could determine if being a sophomore, junior, or senior predicts a different depression level than being a freshman. Consequently, if the variable "junior" is significant in our regression, with a positive coefficient β, this would mean that juniors are significantly more depressive than freshmen. Alternatively, we could have decided to not code "senior", then the coefficients for freshman, sophomore and junior in the regression model would be interpreted as how much more depressive if being a freshman, sophomore, or junior predicts a different depressive level than being a senior.

For the purpose of illustration, the simple regression model with one dummy variable is shown in Fig. 3.2. In the figure, (a) represents regression model with only dummy variable and without regressor. The two groups are parallel. (b) represents the model with dummy variable and regressor x, but two groups are still parallel, (c) represents the model with dummy variable and regressor x. The two groups are not parallel but without crossover. (d) represents the model with dummy variable and regressor x. The two groups are not parallel and with crossover. In situations (c) and (d) we say that there is interaction which means that the response of one group is not always better/higher than the response of the other group by the same magnitude. Situation (c) is quantitative interaction and (d) is qualitative interaction or crossover interaction.

3.17 Collinearity in Multiple Linear Regression

3.17.1 *Collinearity*

What is the collinearity in multiple linear regression? The collinearity refers to the situation in which two or more independent variables in a multiple linear regression model are highly correlated. Let the regression model be $\boldsymbol{y} = \boldsymbol{X} + \varepsilon$ with the design matrix $\boldsymbol{X} = (1, \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k)$. The collinearity occurs if the independent variable $\boldsymbol{x}_i$ is highly linearly correlated to another one or more independent variables $\boldsymbol{x}_{j1}, \boldsymbol{x}_{j2}, \cdots, \boldsymbol{x}_{jk}$. In other words, $\boldsymbol{x}_i$

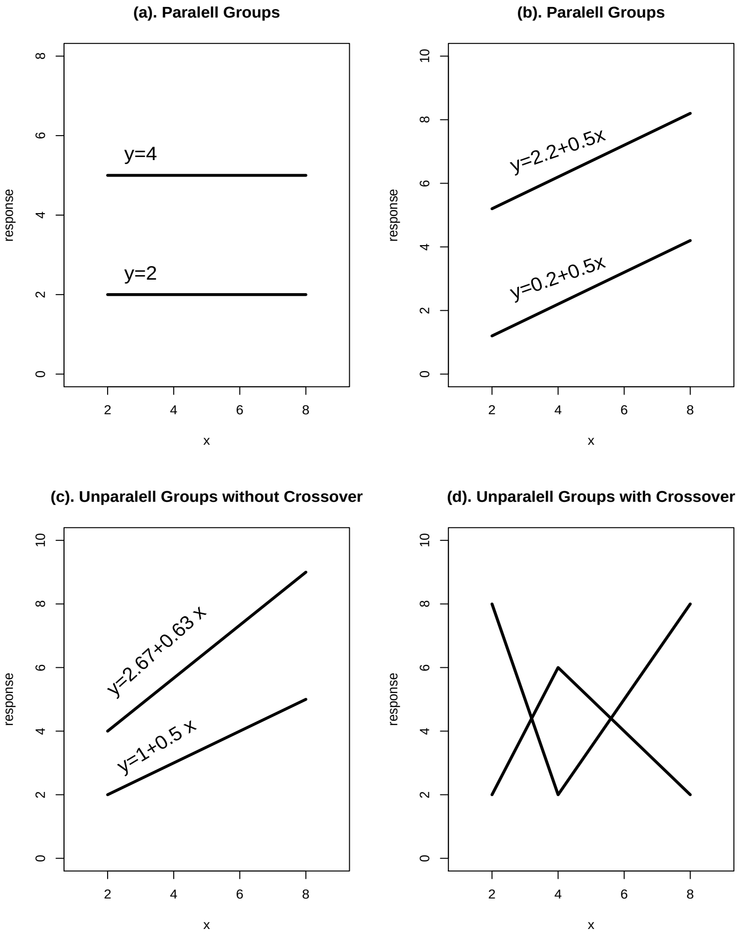

Fig. 3.2 Regression on Dummy Variables

can be almost linearly expressed by one or more other column vectors in $\boldsymbol{X}$. In this situation, the matrix $\boldsymbol{X}^{'}\boldsymbol{X}$ is ill-conditioned or near singular. Although it is not completely singular, its eigenvalues may be close to zero and the eigenvalues of the inverse matrix $(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}$ tend to be very large which may cause instability of the least squares estimates of the regression parameters. If there is a perfect collinearity among column vectors of $\boldsymbol{X}$ then the matrix $\boldsymbol{X}^{'}\boldsymbol{X}$ is not invertible. Therefore, it is problematic to solve for the unique least squares estimators of the regression coefficients from the normal equation. When the column vectors of the design matrix $\boldsymbol{X}$ is highly correlated, then the matrix $\boldsymbol{X}^t\boldsymbol{X}$ becomes ill-conditioned and the least squares estimator become less reliable even though we can find a unique solution of the normal equation. To see this let's look at the following example of two simple data sets (Tables 3.1 and 3.2).

Table 3.1 Two Independent Vectors

x_1	10	10	10	10	15	15	15	15
x_2	10	10	15	15	10	10	15	15

Table 3.2 Two Highly Correlated Vectors

x_1	10.0	11.0	11.9	12.7	13.3	14.2	14.7	15.0
x_2	10.0	11.4	12.2	12.5	13.2	13.9	14.4	15.0

The correlation matrix of the vectors in the first example data is a 2 × 2 identity matrix

$$\boldsymbol{X}^{'}\boldsymbol{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Thus, its inverse matrix is also a 2 × 2 identity matrix. The correlation matrix of the two vectors in the second example data set is

$$\boldsymbol{X}^{'}\boldsymbol{X} = \begin{pmatrix} 1.00000 & 0.99215 \\ 0.99215 & 1.00000 \end{pmatrix}$$

and its inverse matrix is given by

$$(\boldsymbol{X}^{'}\boldsymbol{X})^{-1} = \begin{pmatrix} 63.94 & -63.44 \\ -64.44 & 63.94 \end{pmatrix}.$$

Note that for linear regression, $\text{Var}(\boldsymbol{b}) = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\sigma^2$. For the vectors in the first example data set we have

$$\frac{\text{Var}(b_1)}{\sigma^2} = \frac{\text{Var}(b_2)}{\sigma^2} = 1.$$

For the vectors in the second example data set we have

$$\frac{\text{Var}(b_1)}{\sigma^2} = \frac{\text{Var}(b_2)}{\sigma^2} = 63.94$$

The variances of the regression coefficients are inflated in the example of the second data set. This is because the collinearity of the two vectors in the second data set. The above example is the two extreme cases of the relationship between the two vectors. One is the case where two vectors are orthogonal to each other and the other is the case where two vectors are highly correlated.

Let us further examine the expected Euclidean distance between the least squares estimate $\boldsymbol{b}$ and the true parameter $\boldsymbol{\beta}$, $E(\boldsymbol{b}-\boldsymbol{\beta})^{'}(\boldsymbol{b}-\boldsymbol{\beta})$ when collinearity exists among the column vectors of $\boldsymbol{X}$. First, it is easy to know that $E[(\boldsymbol{b}-\boldsymbol{\beta})^{'}(\boldsymbol{b}-\boldsymbol{\beta})] = E(\boldsymbol{b}^{'}\boldsymbol{b}) - \boldsymbol{\beta}^{'}\boldsymbol{\beta}$. We then calculate $E(\boldsymbol{b}^{'}\boldsymbol{b})$.

$$\begin{aligned}
&E(\boldsymbol{b}^{'}\boldsymbol{b})\\
&= E[(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}]\\
&= E[\boldsymbol{y}^{'}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}]\\
&= (\boldsymbol{X\beta})^{'}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{X\beta} + \sigma^2 tr[\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}]\\
&= \boldsymbol{\beta}^{'}\boldsymbol{X}^{'}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{X\beta} + \sigma^2 tr[\boldsymbol{X}^{'}\boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}]\\
&= \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \sigma^2 tr[(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}]
\end{aligned}$$

Thus, we have

$$E[(\boldsymbol{b}-\boldsymbol{\beta})^{'}(\boldsymbol{b}-\boldsymbol{\beta})] = \sigma^2 tr[(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}].$$

Note that $E[(\boldsymbol{b}-\boldsymbol{\beta})^{'}(\boldsymbol{b}-\boldsymbol{\beta})]$ is the average Euclidean distance measure between the estimate $\boldsymbol{b}$ and the true parameter $\boldsymbol{\beta}$. Assuming that $(\boldsymbol{X}^{'}\boldsymbol{X})$ has k distinct eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_k$, and the corresponding normalized eigenvectors $V = (v_1, v_2, \cdots, v_k)$, we can write

$$V^{'}(\boldsymbol{X}^{'}\boldsymbol{X})V = diag(\lambda_1, \lambda_2, \cdots, \lambda_k).$$

Moreover,

$$tr[V^{'}(\boldsymbol{X}^{'}\boldsymbol{X})V] = tr[VV^{'}(\boldsymbol{X}^{'}\boldsymbol{X})] = tr(\boldsymbol{X}^{'}\boldsymbol{X}) = \sum_{i=1}^{k}\lambda_i.$$

Since the eigenvalues of $(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}$ are $\dfrac{1}{\lambda_1}, \dfrac{1}{\lambda_2}, \cdots, \dfrac{1}{\lambda_k}$ we have

$$E(\boldsymbol{b}^{'}\boldsymbol{b}) = \boldsymbol{\beta}^{'}\boldsymbol{\beta} + \sigma^2 \sum_{i=1}^{k} \frac{1}{\lambda_i},$$

or it can be written as

$$E\Big(\sum_{i=1}^{k} b_i^2\Big) = \sum_{i=1}^{k} \beta_i^2 + \sigma^2 \sum_{i=1}^{k} \frac{1}{\lambda_i}. \tag{3.39}$$

Now it is easy to see that if one of λ is very small, say, $\lambda_i = 0.0001$, then roughly, $\sum_{i=1}^{k} b_i^2$ may over-estimate $\sum_{i=1}^{k} \beta_i^2$ by $1000\sigma^2$ times. The above discussions indicate that if some columns in $\boldsymbol{X}$ are highly correlated with other columns in $\boldsymbol{X}$ then the covariance matrix $(\boldsymbol{X}\boldsymbol{X}^{'})^{-1}\sigma^2$ will have one or more large eigenvalues so that the mean Euclidean distance of $E[(\boldsymbol{b} - \boldsymbol{\beta})^{'}(\boldsymbol{b} - \boldsymbol{\beta})]$ will be inflated. Consequently, this makes the estimation of the regression parameter $\boldsymbol{\beta}$ less reliable. Thus, the collinearity in column vectors of $\boldsymbol{X}$ will have negative impact on the least squares estimates of regression parameters and this need to be examined carefully when doing regression modeling.

How to deal with the collinearity in the regression modeling? One easy way to combat collinearity in multiple regression is to centralize the data. Centralizing the data is to subtract mean of the predictor observations from each observation. If we are not able to produce reliable parameter estimates from the original data set due to collinearity and it is very difficult to judge whether one or more independent variables can be deleted, one possible and quick remedy to combat collinearity in $\boldsymbol{X}$ is to fit the centralized data to the same regression model. This would possibly reduce the degree of collinearity and produce better estimates of regression parameters.

3.17.2 *Variance Inflation*

Collinearity can be checked by simply computing the correlation matrix of the original data $\boldsymbol{X}$. As we have discussed, the variance inflation of the least squares estimator in multiple linear regression is caused by collinearity of the column vectors in $\boldsymbol{X}$. When collinearity exists, the eigenvalues of the covariance matrix $(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\sigma^2$ become extremely large, which causes severe fluctuation in the estimates of regression parameters and makes these estimates less reliable. Variance inflation factor is the measure that can be used to quantify collinearity. The ith variance inflation factor is the scaled version of the multiple correlation coefficient between the ith independent variable and the rest of the independent variables. Specifically, the variance inflation factor for the ith regression coefficient is

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}, \tag{3.40}$$

where R_i^2 is the coefficient of multiple determination of regression produced by regressing the variable x_i against the other independent variables $x_j,\ j \neq i$. Measure of variance inflation is also given as the reciprocal of the above formula. In this case, they are referred to as *tolerances*.

If R_i equals zero (i.e., no correlation between x_i and the remaining independent variables), then VIF_i equals 1. This is the minimum value of variance inflation factor. For the multiple regression model it is recommended looking at the largest VIF value. A VIF value greater than 10 may be an indication of potential collinearity problems. The SAS procedure REG provides information on variance inflation factor and tolerance for each regression coefficient. The following example illustrates how to obtain this information using SAS procedure REG.

Example 3.1. SAS code for detection of collinearity and calculation the variance inflation factor.

```
Data example;
input x1 x2 x3 x4 x5 y;
datalines;
15.57    2463   472.92    18.0   4.45    566.52
44.02    2048  1339.75     9.5   6.92    696.82
20.42    3940   620.25    12.8   4.28   1033.15
18.74    6505   568.33    36.7   3.90   1603.62
49.20    5723  1497.60    35.7   5.50   1611.37
44.92   11520  1365.83    24.0   4.6    1613.27
55.48    5779  1687.00    43.3   5.62   1854.17
59.28    5969  1639.92    46.7   5.15   2160.55
94.39    8461  2872.33    78.7   6.18   2305.58
128.02  20106  3655.08   180.5   6.15   3503.93
96.00   13313  2912.00    60.9   5.88   3571.89
131.42  10771  3921.00   103.7   4.88   3741.40
127.21  15543  3865.67   126.8   5.50   4026.52
```

```
252.90  36194  7684.10   157.7   7.00  10343.81
409.20  34703  12446.33  169.4  10.78  11732.17
463.70  39204  14098.40  331.4   7.05  15414.94
510.22  86533  15524.00  371.6   6.35  18854.45
;
run;

proc reg  data=example corr alpha=0.05;
          model y=x1 x2 x3 x4 x5/tol vif collin;
run;

*Fit the regression model after deleting variable X1;
proc reg data=example corr alpha=0.05; ;
     model y=x2 x3 x4 x5/tol vif collin;
run;
```

The keyword TOL requests tolerance values for the estimates, VIF gives the variance inflation factors with the parameter estimates, and COLLIN requests a detailed analysis of collinearity among regressors. Variance inflation (VIF) is the reciprocal of tolerance (TOL). The above SAS procedures produce the following Table 3.3. The table shows that variables x_1 and x_3 are highly correlated. Due to this high correlation the variance inflation for both the variables x_1 and x_3 are rather significant and it can be found in Table 3.4.

We then delete variable x_1 and recalculate the correlation matrix. It can be seen that the variance inflations for all independent variables become much smaller after deleting x_1. The results of the correlation matrix and variance inflation are presented in Tables 3.5 and 3.6.

The least squares estimates in the regression model including the independent variables x_2, x_3, x_4 and x_5 behave much better than the model including all independent variables. The collinearity is eliminated by deleting one independent variable x_1 in this example.

3.18 Linear Model in Centered Form

The linear model can be rewritten in terms of centered x's as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \\ &= \alpha + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \cdots + \beta_k(x_{ik} - \bar{x}_k) + \varepsilon_i \quad (3.41) \end{aligned}$$

Table 3.3 Correlation Matrix for Variables $x_1, x_2, \cdots, x_5$

Variable	x_1	x_2	x_3	x_4	x_5	y
x1	1.0000	0.9074	0.9999	0.9357	0.6712	0.9856
x2	0.9074	1.0000	0.9071	0.9105	0.4466	0.9452
x3	0.9999	0.9071	1.0000	0.9332	0.6711	0.9860
x4	0.9357	0.9105	0.9332	1.0000	0.4629	0.9404
x5	0.6712	0.4466	0.6711	0.4629	1.0000	0.5786
y	0.9856	0.9452	0.9860	0.9404	0.5786	1.0000

Table 3.4 Parameter Estimates and Variance Inflation

Variable	Parameter	STD	t value	$P > \|t\|$	Tolerance	Inflation
Intercept	1962.95	1071.36	1.83	0.094		0
x1	-15.85	97.66	-0.16	0.874	0.0001042	9597.57
x2	0.06	0.02	2.63	0.023	0.12594	7.94
x3	1.59	3.09	0.51	0.617	0.000112	8933.09
x4	-4.23	7.18	-0.59	0.569	0.04293	23.29
x5	-394.31	209.64	-1.88	0.087	0.23365	4.28

Table 3.5 Correlation Matrix after Deleting Variable x_1

Variable	x_2	x_3	x_4	x_5	y
x2	1.0000	0.9071	0.9105	0.4466	0.9452
x3	0.9071	1.0000	0.9332	0.6711	0.9860
x4	0.9105	0.9332	1.0000	0.4629	0.9404
x5	0.4466	0.6711	0.4629	1.0000	0.5786
y	0.9452	0.9860	0.9404	0.5786	1.0000

Table 3.6 Variance Inflation after Deleting x_1

variable	parameter	std	t value	$P > \|t\|$	tolerance	inflation
intercept	2032.19	942.075	2.16	0.0520	0	
x2	0.056	0.020	2.75	0.0175	0.126	7.926
x3	1.088	0.153	7.10	$< .0001$	0.042	23.927
x4	-5.00	5.081	-0.98	0.3441	0.079	12.706
x5	-410.083	178.078	-2.30	0.0400	0.298	3.361

for $i = 1, \ldots, n$, where

$$\alpha = \beta_0 + \beta_1 \bar{x}_1 + \cdots + \beta_k \bar{x}_k$$

or

$$\beta_0 = \alpha - (\beta_1 \bar{x}_1 + \cdots + \beta_k \bar{x}_k) = \alpha - \bar{\mathbf{x}}' \boldsymbol{\beta}_1; \tag{3.42}$$

$\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k)'$; $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_k)'$; and $\bar{x}_j$ denotes the sample average of x_{ij}'s for $j = 1, \ldots, k$. In the centered form, Y is regressed on centered X's, in which case the slope parameters in $\boldsymbol{\beta}_1$ remain the same. This centered form sometimes brings convenience in derivations of estimators of the linear models. Also, one can try the regression model in centered form when collinearity is observed among the independent variables and independent variables are difficult to be eliminated. Expressed in matrix form, model (3.41) becomes

$$\mathbf{y} = (\mathbf{j}, \mathbf{X}_c) \begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{\varepsilon}, \tag{3.43}$$

where

$$\mathbf{X}_c = \left(\mathbf{I} - \frac{1}{n} \mathbf{J} \right) \mathbf{X}_1 = (x_{ij} - \bar{x}_j)\,; \tag{3.44}$$

and $\mathbf{X}_1 = (x_{ij})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, k$. Here matrix $\mathbf{X}_1$ is the sub-matrix of $\mathbf{X}$ after removing the first column of all 1's.

The matrix $\mathbf{C} = \mathbf{I} - 1/n \cdot \mathbf{J}$ is called the *centering matrix*, where $\mathbf{J} = \mathbf{j}\mathbf{j}'$ is an $n \times n$ matrix with all elements being 1. A geometric look at the centering matrix shows that

$$\begin{aligned} \mathbf{C} &= \mathbf{I} - \frac{1}{n} \cdot \mathbf{J} = \mathbf{I} - \mathbf{j}(\mathbf{j}'\mathbf{j})^{-1}\mathbf{j}', \quad \text{noting } \mathbf{j}'\mathbf{j} = n \\ &= \mathbf{I} - \mathbf{P}_{\mathcal{W}} = \mathbf{P}_{\mathcal{W}^\perp}, \end{aligned}$$

where $\mathcal{W} = C(\mathbf{j})$ denotes the subspace spanned by $\mathbf{j}$; $\mathcal{W}^\perp$ is the subspace perpendicular to $\mathcal{W}$; and $\mathbf{P}_{\mathcal{W}}$ and $\mathbf{P}_{\mathcal{W}^\perp}$ are their respective projection matrices. Namely, matrix $\mathbf{C}$ is the project matrix on the subspace that is perpendicular to the subspace spanned by $\mathbf{j}$. It follows immediately that

$$(\mathbf{I} - \frac{1}{n} \cdot \mathbf{J})\mathbf{j} = \mathbf{0} \text{ and } \mathbf{j}'\mathbf{X}_c = \mathbf{0} \tag{3.45}$$

Using (3.45), the least squared estimators of $(\alpha, \boldsymbol{\beta}_1)$ are given by,

$$\begin{aligned} \begin{pmatrix} \hat{\alpha} \\ \widehat{\boldsymbol{\beta}}_1 \end{pmatrix} &= \{(\mathbf{j},\ \mathbf{X}_c)'(\mathbf{j},\ \mathbf{X}_c)\}^{-1} (\mathbf{j},\ \mathbf{X}_c)'\mathbf{y} = \begin{pmatrix} n & \mathbf{0}' \\ \mathbf{0} & \mathbf{X}_c'\mathbf{X}_c \end{pmatrix}^{-1} \begin{pmatrix} n\bar{y} \\ \mathbf{X}_c'\mathbf{y} \end{pmatrix} \\ &= \begin{pmatrix} 1/n & \mathbf{0}' \\ \mathbf{0} & (\mathbf{X}_c'\mathbf{X}_c)^{-1} \end{pmatrix} \begin{pmatrix} n\bar{y} \\ \mathbf{X}_c'\mathbf{y} \end{pmatrix} \\ &= \begin{pmatrix} \bar{y} \\ (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y} \end{pmatrix}. \end{aligned}$$

Thus, $\widehat{\boldsymbol{\beta}}_1$ is the same as in the ordinary least squares estimator $\widehat{\boldsymbol{\beta}}_1$ and

$$\hat{\beta}_0 = \bar{y} - \bar{\mathbf{x}}'\widehat{\boldsymbol{\beta}}_1 \tag{3.46}$$

in view of (3.42) and uniqueness of LSE.

Using the centered form, many interesting properties of the least squares estimation can be easily obtained. First, the LS fitted regression plane satisfies

$$y - \hat{\alpha} = y - \bar{y} = (\mathbf{x} - \bar{\mathbf{x}})'\widehat{\boldsymbol{\beta}}_1$$

and hence must pass through the center of the data $(\bar{\mathbf{x}}, \bar{y})$.

Denote $\mathcal{V}_c = C(\mathbf{X}_c)$. Since $\mathcal{W} = C(\mathbf{j}) \perp \mathcal{V}_c$ using (3.45),

$$\mathcal{V} = C(\mathbf{X}) = \mathcal{W} \oplus \mathcal{V}_c.$$

The vector fitted values is

$$\widehat{\mathbf{y}} = \mathbf{P}_{\mathcal{V}}\mathbf{y} = \mathbf{P}_{\mathcal{W}}\mathbf{y} + \mathbf{P}_{\mathcal{V}_c}\mathbf{y} = \bar{y}\mathbf{j} + \mathbf{X}_c(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y} = \bar{y}\mathbf{j} + \mathbf{X}_c\widehat{\boldsymbol{\beta}}_1 \tag{3.47}$$

and the residual vector is

$$\boldsymbol{e} = (\mathbf{I} - \mathbf{P}_{\mathcal{W}} - \mathbf{P}_{\mathcal{V}_c})\mathbf{y} = (\mathbf{P}_{\mathcal{W}^\perp} - \mathbf{P}_{\mathcal{V}_c})\mathbf{y}. \tag{3.48}$$

Consider the sum of squared error (SSE), which becomes

$$\begin{aligned} \text{SSE} &= \parallel \mathbf{y} - \widehat{\mathbf{y}} \parallel^2 = \boldsymbol{e}'\boldsymbol{e} \\ &= \mathbf{y}'(\mathbf{P}_{\mathcal{W}^\perp} - \mathbf{P}_{\mathcal{V}_c})\mathbf{y} = \mathbf{y}'\mathbf{P}_{\mathcal{W}^\perp}\mathbf{y} - \mathbf{y}'\mathbf{P}_{\mathcal{V}_c}\mathbf{y} \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \widehat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y} = \text{SST} - \widehat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}. \end{aligned} \tag{3.49}$$

Namely, the sum of squares regression (SSR) is SSR $= \widehat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}$.
The leverage $h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ can also be reexpressed for better interpretation using the centered form. Letting $\mathbf{x}_{1i} = (x_{i1}, x_{2i}, \ldots, x_{ik})'$,

$$\begin{aligned} h_i &= (1,\ \mathbf{x}_{1i}' - \bar{\mathbf{x}}')\left\{(\mathbf{j},\ \mathbf{X}_c)'(\mathbf{j},\ \mathbf{X}_c)\right\}^{-1}\begin{pmatrix} 1 \\ \mathbf{x}_{1i} - \bar{\mathbf{x}} \end{pmatrix} \\ &= (1,\ \mathbf{x}_{1i}' - \bar{\mathbf{x}}')\begin{pmatrix} 1/n & \mathbf{0}' \\ \mathbf{0} & (\mathbf{X}_c'\mathbf{X}_c)^{-1} \end{pmatrix}\begin{pmatrix} 1 \\ \mathbf{x}_{1i} - \bar{\mathbf{x}} \end{pmatrix} \\ &= \frac{1}{n} + (\mathbf{x}_{1i} - \bar{\mathbf{x}})'(\mathbf{X}_c\mathbf{X}_c)^{-1}(\mathbf{x}_{1i} - \bar{\mathbf{x}}). \end{aligned} \tag{3.50}$$

Note that

$$\mathbf{X}_c\mathbf{X}_c = (n-1)\mathbf{S}_{xx}, \tag{3.51}$$

where

$$\mathbf{S}_{xx} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1k} \\ s_{21} & s_2^2 & \cdots & s_{2k} \\ \vdots & \vdots & & \vdots \\ s_{k1} & s_{k2} & \cdots & s_k^2 \end{pmatrix} \text{ with } \begin{cases} s_j^2 &= \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \\ s_{jj'} &= \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) \end{cases}$$

is the sample variance-covariance matrix for $\mathbf{x}$ vectors. Therefore, h_i in (3.50) is

$$h_i = \frac{1}{n} + \frac{(\mathbf{x}_{1i} - \bar{\mathbf{x}})'\mathbf{S}_{xx}^{-1}(\mathbf{x}_{1i} - \bar{\mathbf{x}})}{n-1}. \tag{3.52}$$

Clearly, the term $(\mathbf{x}_{1i} - \bar{\mathbf{x}})'\mathbf{S}_{xx}^{-1}(\mathbf{x}_{1i} - \bar{\mathbf{x}})$ gives the Mahalanobis distance between $\mathbf{x}_{1i}$ and the center of the data $\bar{\mathbf{x}}$, which renders h_i an important diagnostic measure for assessing how outlying an observation is in terms of its predictor values.

Furthermore, both $\hat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_1$ can be expressed in terms of the sample variances and covariances. Let $\mathbf{s}_{yx}$ denote the covariance vector between Y and X_j's. Namely,

$$\mathbf{s}_{yx} = (s_{y1}, s_{y2}, \ldots, s_{yk})', \tag{3.53}$$

where

$$s_{yj} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j) \cdot (y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j) \cdot y_i}{n-1}.$$

It can be easily seen that

$$(n-1) \cdot \mathbf{s}_{yx} = \mathbf{X}'\mathbf{y}. \tag{3.54}$$

Using equations (3.51) and (3.54), we have

$$\widehat{\boldsymbol{\beta}}_1 = \left(\frac{\mathbf{X}_c'\mathbf{X}_c}{n-1}\right)^{-1} \frac{\mathbf{X}_c'\mathbf{y}}{n-1} = \mathbf{S}_{xx}^{-1}\mathbf{s}_{yx} \tag{3.55}$$

and

$$\hat{\beta}_0 = \bar{y} - \widehat{\boldsymbol{\beta}}_1'\bar{\mathbf{x}} = \bar{y} - \mathbf{s}_{yx}'\mathbf{S}_{xx}^{-1}\bar{\mathbf{x}}. \tag{3.56}$$

The above forms are now analogous to those formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$ in simple linear regression.

Besides, the coefficient of determination R^2 can also be expressed in terms of $\mathbf{S}_{xx}$ and $\mathbf{s}_{yx}$ as below

$$\begin{aligned} R^2 = \frac{\text{SSR}}{SST} &= \frac{\widehat{\boldsymbol{\beta}}_1' \mathbf{X}_c' \mathbf{X}_c \widehat{\boldsymbol{\beta}}_1}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\mathbf{s}_{yx}' \mathbf{S}_{xx}^{-1} (n-1) \mathbf{S}_{xx} \mathbf{S}_{xx}^{-1} \mathbf{s}_{yx}}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\mathbf{s}_{yx}' \, \mathbf{S}_{xx}^{-1} \, \mathbf{s}_{yx}}{s_y^2}. \end{aligned} \tag{3.57}$$

3.19 Numerical Computation of LSE via QR Decomposition

According to earlier derivation, the least squares estimator $\widehat{\boldsymbol{\beta}}$ is obtained by solving the normal equations

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}. \tag{3.58}$$

Nevertheless, the approach is not very computationally attractive because it can be difficult to form matrices in (3.58) to a great numerical accuracy. Instead, computation of LSE, as well as many other related quantities, is carried out through QR decomposition of the design matrix $\mathbf{X}$. The basic idea of this approach utilizes a successive orthogonalization process on the predictors to form an orthogonal basis for the linear space $\mathcal{V} = C(\mathbf{X})$.

3.19.1 *Orthogonalization*

To motivate, we first consider the simple regression (Y versus X) with design matrix $\mathbf{X} = (\mathbf{j}, \mathbf{x})$. The LSE of β_1 is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\langle \mathbf{x} - \bar{x}\mathbf{j}, \mathbf{y} \rangle}{\langle \mathbf{x} - \bar{x}\mathbf{j}, \mathbf{x} - \bar{x}\mathbf{j} \rangle},$$

where $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^t \mathbf{y}$ denotes the inner product between $\mathbf{x}$ and $\mathbf{y}$. The above estimate $\hat{\beta}_1$ can be obtained in two steps, either applying a simple linear regression without intercept. In step 1, regress $\mathbf{x}$ on $\mathbf{j}$ without intercept and obtain the residual $\boldsymbol{e} = \mathbf{x} - \bar{x}\mathbf{j}$; and in step 2, regress $\mathbf{y}$ on the residual $\boldsymbol{e}$ without intercept to produce $\hat{\beta}_1$.

Note that regressing $\mathbf{u}$ on $\mathbf{v}$ without intercept by fitting model

$$u_i = \gamma v_i + \varepsilon_i$$

gives

$$\hat{\gamma} = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \quad \text{and residual vector} \quad \boldsymbol{e} = \mathbf{u} - \hat{\gamma}\mathbf{v}, \tag{3.59}$$

which is exactly the step in linear algebra taken to orthogonalize one vector $\mathbf{u}$ with respect to another vector $\mathbf{v}$. The key point is to ensure residual vector $\boldsymbol{e}$ to be orthogonal to $\mathbf{v}$, i.e., $\boldsymbol{e} \perp \mathbf{v}$ or $\langle \boldsymbol{e}, \mathbf{v} \rangle = 0$.

Orthogonality often provides great convenience and efficiency in designed experiments. It is easy to show, for example, that if the k predictors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ in a multiple linear regression model are orthogonal to each other, then the LSE of the j-th slope equals

$$\hat{\beta}_j = \frac{\langle \mathbf{x}_j, \mathbf{y} \rangle}{\langle \mathbf{x}_j, \mathbf{x}_j \rangle},$$

which is the same as the slope estimator obtained in a simple linear regression model that regresses $\mathbf{y}$ on $\mathbf{x}_j$. This implies that orthogonal predictors have no confounding effect on each other at all.

In the two-step approach, the subspace $C(\mathbf{j}, \mathbf{x})$ spanned by $(\mathbf{j}, \mathbf{x})$ is the same as the subspace spanned by the orthogonal basis $(\mathbf{j}, \boldsymbol{e})$. This idea can be generalized to multiple linear regression, which leads to the algorithm outlined below. This is the well-known Gram-Schmidt procedure for constructing an orthogonal basis from an arbitrary basis. Given a design matrix $\mathbf{X} = (\mathbf{x}_0 = \mathbf{j}, \mathbf{x}_1, \ldots, \mathbf{x}_k)$ with columns $\mathbf{x}_j$, the result of the algorithm is an orthogonal basis $(\boldsymbol{e}_0, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_k)$ for the column subspace of $\mathbf{X}$, $\mathcal{V} = C(\mathbf{X})$.

Algorithm 9.1: Gram-Schmidt Algorithm for Successive Orthogonalization.

- Set $\boldsymbol{e}_0 = \mathbf{j}$;
- Compute $\gamma_{01} = \langle \mathbf{x}_1, \boldsymbol{e}_0 \rangle / \langle \boldsymbol{e}_0, \boldsymbol{e}_0 \rangle$ and $\boldsymbol{e}_1 = \mathbf{x}_1 - \gamma_{01}\, \boldsymbol{e}_0$;
- Compute $\gamma_{02} = \langle \mathbf{x}_2, \boldsymbol{e}_0 \rangle / \langle \boldsymbol{e}_0, \boldsymbol{e}_0 \rangle$ and $\gamma_{12} = \langle \mathbf{x}_2, \boldsymbol{e}_1 \rangle / \langle \boldsymbol{e}_1, \boldsymbol{e}_1 \rangle$ and obtain $\boldsymbol{e}_2 = \mathbf{x}_2 - (\gamma_{02}\, \boldsymbol{e}_0 + \gamma_{12}\, \boldsymbol{e}_1)$.

$\vdots$

- Continue the process up to $\mathbf{x}_k$, which involves computing $(\gamma_{0k}, \gamma_{1k}, \ldots, \gamma_{(k-1)k})$ with $\gamma_{jk} = \langle \mathbf{x}_k, \boldsymbol{e}_j \rangle / \langle \boldsymbol{e}_j, \boldsymbol{e}_j \rangle$ for $j = 0, 1, \ldots, (k-1)$ and then obtaining $\boldsymbol{e}_k = \mathbf{x}_k - (\gamma_{0k}\, \boldsymbol{e}_0 + \gamma_{1k}\, \boldsymbol{e}_1 + \cdots + \gamma_{(k-1)k}\, \boldsymbol{e}_{k-1})$.

Note that $\boldsymbol{e}_j \perp \boldsymbol{e}_{j'}$ for $j \neq j'$. It is interesting and insightful to take a few observations, as listed below. First, the slope estimate obtained by regressing $\mathbf{y}$ on $\boldsymbol{e}_k$ without intercept is the same as the LS slope estimate for $\mathbf{x}_k$ in multiple linear model of $\mathbf{y}$ versus $(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k)$. That is,

$$\hat{\beta}_k = \frac{\langle \mathbf{y}, \boldsymbol{e}_k \rangle}{\langle \boldsymbol{e}_k, \boldsymbol{e}_k \rangle} = \frac{\langle \mathbf{y}, \boldsymbol{e}_k \rangle}{\| \boldsymbol{e}_k \|^2}. \tag{3.60}$$

This can be verified by using the fact that $\boldsymbol{e}_j$'s form an orthogonal basis for the column space of $\mathbf{X}$ and $\mathbf{x}_k$ is only involved in $\boldsymbol{e}_k = \mathbf{x}_k - \sum_{j=0}^{k-1} \gamma_{jk} \boldsymbol{e}_j$, with coefficient 1.

Secondly, since $(\boldsymbol{e}_0, \boldsymbol{e}_1, \ldots \boldsymbol{e}_{k-1})$ spans the same subspace as $(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{k-1})$ does, the residual vector $\boldsymbol{e}_k$ is identical to the residual vector obtained by regressing $\mathbf{x}_k$ versus $(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{k-1})$. This is the result that motivates the partial regression plots, in which the residuals obtained from regressing $\mathbf{y}$ on $(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{k-1})$ are plotted versus $\boldsymbol{e}_k$.

Thirdly, the same results clearly hold for any one of the predictors if one rearranges it to the last position. The general conclusion is that the j-th slope $\hat{\beta}_j$ can be obtained by fitting a simple linear regression of $\mathbf{y}$ on the residuals obtained from regressing $\mathbf{x}_j$ on other predictors $(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_k)$. Thus, $\hat{\beta}_j$ can be interpreted as the additional contribution of $\mathbf{x}_j$ on $\mathbf{y}$ after $\mathbf{x}_j$ has been adjusted for other predictors. Furthermore, from (3.60) the variance of $\hat{\beta}_k$ is

$$\text{Var}(\hat{\beta}_k) = \frac{\sigma^2}{\| \boldsymbol{e}_k \|^2}. \tag{3.61}$$

In other words, $\boldsymbol{e}_k$, which represents how much of $\mathbf{x}_p$ is unexplained by other predictors, plays an important role in estimating β_k.

Fourthly, if $\mathbf{x}_p$ is highly correlation with some of the other predictors, a situation to which multicolinearity is referred, then the residual vector $\boldsymbol{e}_k$ will be close to zero in length $|\boldsymbol{e}|$. From (3.60), the estimate $\hat{\beta}_k$ would be very unstable. From (3.61), the precision of the estimate would be poor as well. The effect due to multicolinearity is clearly true for all predictors in the correlated predictors.

3.19.2 *QR Decomposition and LSE*

The Gram-Schmidt algorithm can be represented in matrix form. It can be easily verified that

$$\gamma_{jl} = \frac{\langle \mathbf{x}_l, \boldsymbol{e}_j \rangle}{\langle \boldsymbol{e}_j, \boldsymbol{e}_j \rangle} = \begin{cases} 1 & \text{if } \; j = l \\ 0 & \text{if } \; j < l \end{cases} \tag{3.62}$$

for $j, l = 0, 1, \ldots, k$. Denoting

$$\mathbf{\Gamma} = (\gamma_{jl}) = \begin{pmatrix} 1 & \gamma_{01} & \gamma_{02} & \cdots & \gamma_{0(k-1)} & \gamma_{0k} \\ 0 & 1 & \gamma_{12} & \cdots & \gamma_{1(k-1)} & \gamma_{1k} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \gamma_{(k-1)k} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{E} = (\boldsymbol{e}_0, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_k),$$

we have

$$\mathbf{X} = \mathbf{E\Gamma} = (\mathbf{ED}^{-1})(\mathbf{D\Gamma}) = \mathbf{QR}, \tag{3.63}$$

where $\mathbf{D} = \text{diag}(d_{jj})$ with $d_{jj} =\parallel \boldsymbol{e}_{j-1} \parallel$ for $j = 1, \ldots, (k+1)$. The form given in (3.63) is the so-called QR decomposition of $\mathbf{X}$. The matrix

$$\mathbf{Q} = \mathbf{ED}^{-1} \tag{3.64}$$

is an orthogonal matrix of dimension $n \times (k+1)$ satisfying $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ and the matrix

$$\mathbf{R} = \mathbf{D\Gamma} \tag{3.65}$$

is a $(k+1) \times (k+1)$ upper triangular matrix.

Using the decomposition in (3.63), the normal equations becomes

$$\begin{aligned} \mathbf{X}^t\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}^t\mathbf{y} \\ \Longrightarrow \mathbf{R}'\mathbf{Q}^t\mathbf{QR}\boldsymbol{\beta} &= \mathbf{R}'\mathbf{Q}'\mathbf{y} \\ \Longrightarrow \qquad \mathbf{R}\boldsymbol{\beta} &= \mathbf{Q}\mathbf{y}, \end{aligned} \tag{3.66}$$

which are easy to solve since $\mathbf{R}$ is upper triangular. This leads to the LSE

$$\widehat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}'\mathbf{y}. \tag{3.67}$$

Its variance-covariance matrix can be expressed as

$$\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \cdot (\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2 \cdot (\mathbf{R}^t\mathbf{R})^{-1} = \sigma^2 \cdot \mathbf{R}^{-1}(\mathbf{R}^t)^{-1}. \tag{3.68}$$

To compute, $\mathbf{R}^{-1}$ is needed. Since $\mathbf{R}$ is upper triangular, $\mathbf{R}^{-1} = \mathbf{W}$ can be easily obtained with back-substitution in the system of linear equations

$$\mathbf{R}\,\mathbf{W} = \mathbf{I}. \tag{3.69}$$

Various other desired quantities in linear regression including the F test statistic for linear hypotheses can also be computed using the QR decomposition. For example, the predicted vector is

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{Q}\mathbf{Q}^t\mathbf{y},$$

the residual vector is

$$\boldsymbol{e} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^t)\mathbf{y}$$

and the sum of squared errors (SSE) is

$$\text{SSE} = \boldsymbol{e}^t\boldsymbol{e} = \mathbf{y}^t(\mathbf{I} - \mathbf{Q}\mathbf{Q}^t)\mathbf{y} =\parallel \mathbf{y} \parallel^2 - \parallel \mathbf{Q}^t\mathbf{y} \parallel^2 .$$

3.20 Analysis of Regression Residual

3.20.1 *Purpose of the Residual Analysis*

Definition 3.5. The residual of the linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is defined as the difference between observed response variable $\boldsymbol{y}$ and the fitted value $\hat{\boldsymbol{y}}$, i.e., $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$.

The regression error term $\boldsymbol{\varepsilon}$ is unobservable and the residual is observable. Residual is an important measurement of how close the calculated response from the fitted regression model to the observed response. The purposes of the residual analysis are to detect model mis-specification and to verify model assumptions. Residuals can be used to estimate the error term in regression model, and the empirical distribution of residuals can be utilized to check the normality assumption of the error term (QQ plot), equal variance assumption, model over-fitting, model under-fitting, and outlier detection. Overall, residual analysis is useful for assessing a regression model.

Simple statistical properties of the regression residual can be discussed. The ith residual of the linear regression model can be written as

$$e_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i b = y_i - \boldsymbol{x}_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}.$$

Regression residual can be expressed in a vector form

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y} = (I - \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'})\boldsymbol{y} = (I - H)\boldsymbol{y}, \quad (3.70)$$

where $H = \boldsymbol{X}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}$ is called the HAT matrix. Note that $I - H$ is symmetric and idempotent, i.e., $I - H = (I - H)^2$. The covariance matrix of the residual $\boldsymbol{e}$ is given by:

$$\text{Cov}(\boldsymbol{e}) = (I - H)\text{Var}(\boldsymbol{y})(I - H)^{'} = (I - H)\sigma^2.$$

Denote the hat matrix $H = (h_{ij})$ we have

$$\text{Var}(e_i) = (1 - h_{ii})\sigma^2$$

and

$$\text{Cov}(e_i, e_j) = -h_{ij}\sigma^2.$$

The HAT matrix contains useful information for detecting outliers and identifying influential observations.

Table 3.7 United States Population Data (in Millions)

Year	Population	Year	Population	Year	Population	Year	Population
1790	3.929	1800	5.308	1810	7.239	1820	9.638
1830	12.866	1840	17.069	1850	23.191	1860	31.443
1870	39.818	1880	50.155	1890	62.947	1900	75.994
1910	91.972	1920	105.710	1930	122.775	1940	131.669
1950	151.325	1960	179.323	1970	203.211		

3.20.2 *Residual Plot*

A plot of residuals e_i's against the fitted values $\hat{y}_i$'s is residual plot, which is a simple and convenient tool for regression model diagnosis. The residuals evenly distributed on both sides of $y = 0$ imply that the assumptions $E(\varepsilon) = 0$ and constant variance $\text{Var}(\varepsilon_i) = \sigma^2$ are appropriate. A curvature appearance in residual plot implies that some higher order terms in regression model may be missing. A funnel shape of residual plot indicates heterogeneous variance and violation of model assumption $\text{Var}(\varepsilon) = \sigma^2$. In addition, periodical and curvature residuals may indicate that the possible regression model may be piecewise and some higher order terms in the model may be missing. The following Fig. 3.3 illustrate different situations of the residuals in regression model. Figure (a) displays residuals evenly distributed about 0, (b) shows residuals with uneven variances, (c) displays residuals with curvature pattern, and (d) displays periodic and curvature residuals. The classical regression model assumptions $E(\boldsymbol{\varepsilon}) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$ are satisfied only when residuals are evenly distributed about 0. The other residual plots imply some deviations from the classical regression model assumptions. When model assumptions are violated the model is no longer valid and statistical inference based on model is not reliable anymore.

We now discuss how to use residual plot to improve regression model. The illustrative example is the regression model of the populations of the United States from Year 1790 to Year 1970. We will show how to improve the regression model based on residual diagnosis. The population data (in millions) are presented in Table 3.7.

Example 3.2. First, we fit the data to the simple linear regression model

$$\text{Population} = \beta_0 + \beta_1 \text{Year} + \varepsilon.$$

The estimates of the model parameters are presented in Table 3.8. We then compute the residuals and plot the residuals against the fitted values

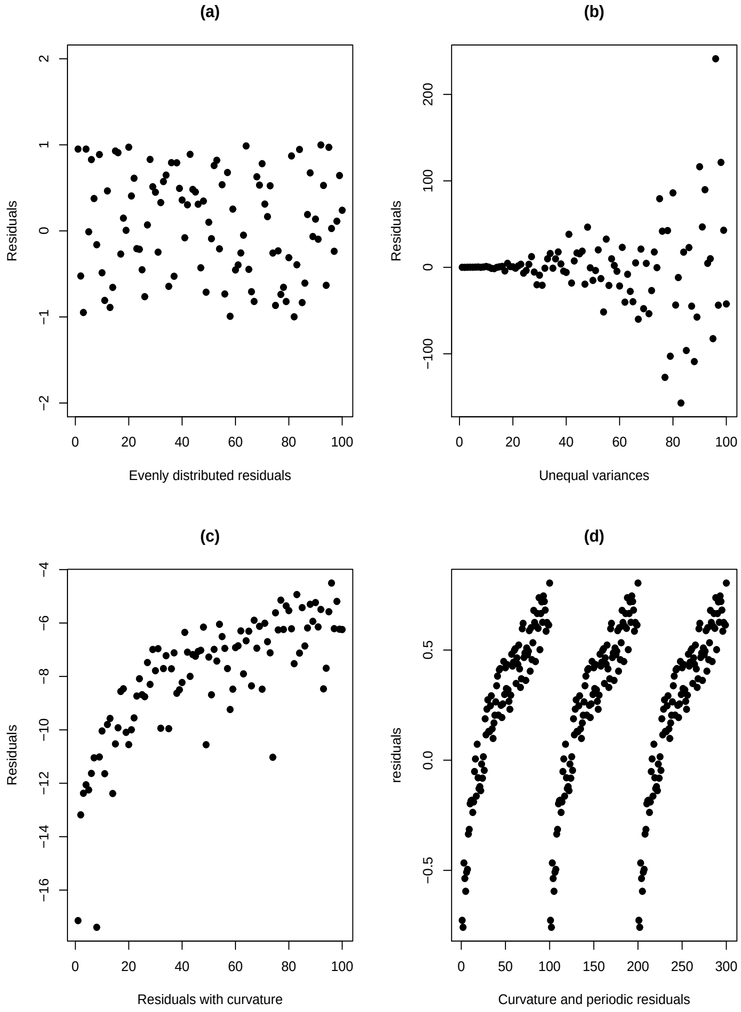

Fig. 3.3 Various Shapes of Residual Plots

$\hat{y}$ for regression model Population $= \beta_0 + \beta_1 \text{Year} + \varepsilon$. The residual plot is presented in Fig. 3.4 (b). The curvature appearance of the residual plot implies that the proposed regression model may be under-fitted. i.e., some necessary higher order terms in the regression model may be missing.

Table 3.8 Parameter Estimates for Model Population=Year

MODEL	TYPE	DEPVAR	RMSE	Intercept	year
MODEL1	PARMS	population	18.1275	-1958.37	1.0788
MODEL1	STDERR	population	18.1275	142.80	0.0759
MODEL1	T	population	18.1275	-13.71	14.2082
MODEL1	PVALUE	population	18.1275	0.00	0.0000
MODEL1	L95B	population	18.1275	-2259.66	0.9186
MODEL1	U95B	population	18.1275	-1657.08	1.2390

Table 3.9 Parameter Estimates for Model Population=Year+Year^2

TYPE	DEPVAR	RMSE	Intercept	Year	Year^2
PARMS	population	2.78102	20450.43	-22.7806	0.0063
STDERR	population	2.78102	843.48	0.8978	0.0002
T	population	2.78102	24.25	-25.3724	26.5762
PVALUE	population	2.78102	0.00	0.0000	0.0000
L95B	population	2.78102	18662.35	-24.684	0.0058
U95B	population	2.78102	22238.52	-20.8773	0.0069

The curvature of a quadratic appearance in the residual plot suggests that a quadratic term in the model may be missing. We then add a term $Year^2$ into the model and fit the data to the following regression model:

$$\text{Population} = \beta_0 + \beta_1\text{Year} + \beta_2\text{Year}^2 + \varepsilon.$$

The estimates of the model parameters are presented in Table 3.9. The residual plot of above regression model is presented in Fig. 3.4 (c) and the shape of the residual plot is clearly better than the residual plot Fig. 3.4 (b), since residuals become more evenly distributed on both sides of $y = 0$. If we take a closer look at the residual plot, we still observe that two residuals, which are at years 1940 and 1950, are far from the line $y = 0$. We know that in the history these are the years during the World War II. We think there might be a shift of populations due to the war. So we try to add a dummy variable z into the model. The dummy variable z takes value 1 at years 1940 and 1950, and 0 elsewhere. The regression model with mean shift term z can be written as

$$\text{Population} = \beta_0 + \beta_1\text{Year} + \beta_2\text{Year}^2 + \beta_3 z + \varepsilon.$$

We then fit the US population data to the above model. The estimates of the model parameters are presented in Table 3.10. The residual plot for

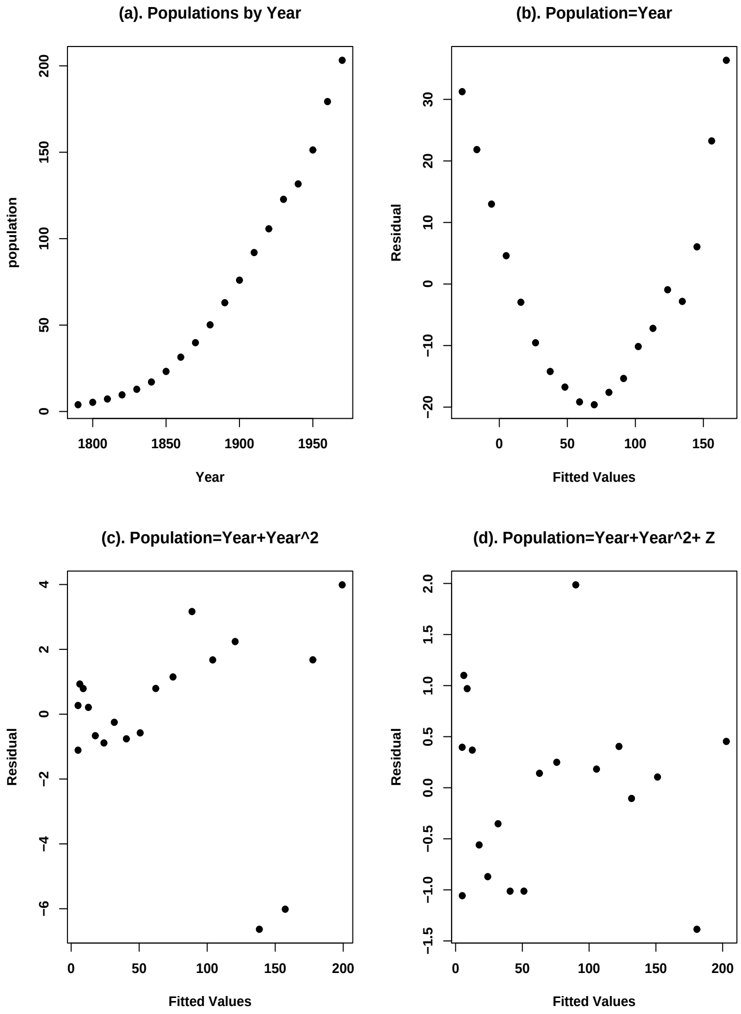

Fig. 3.4 Residual Plots of the Regression Model of the US Population

this model is presented in Fig 3.4 (d) and it is clearly improved, since the residuals are much more evenly distributed on both sides of $y = 0$, including the residuals at Years 1940 and 1950.

The SAS program for generating analysis results above is provided below for illustrative purpose.

Table 3.10 Parameter Estimates for Regression Model Population=β_0 +β_1 Year+ β_2 Year^2+z

TYPE	DEPVAR	RMSE	Intercept	Year	Year^2	z
PARMS	population	0.93741	20982.75	-23.3664	0.0065	-8.7415
STDERR	population	0.93741	288.25	0.3071	0.0001	0.7793
T	population	0.93741	72.79	-76.0838	79.5883	-11.2170
PVALUE	population	0.93741	0.00	0.0000	0.0000	0.0000
L95B	population	0.93741	20368.37	-24.0210	0.0063	-10.4026
U95B	population	0.93741	21597.14	-22.7118	0.0067	-7.0805

```
data pop; set pop;
yyear=year*year;
if year in (1940, 1950) then z=1;
else z=0;
run;

proc reg data=pop outest=out1 tableout;
     model population=year;
     output out=out2
     p=yhat r=yresid student=sresid;
run;

proc reg data=pop outest=out3 tableout;
     model population=year yyear;
     output out=out4
     p=yhat r=yresid student=sresid;
run;

proc reg data=pop outest=out5 tableout;
     model population=year yyear z;
     output out=out6
     p=yhat r=yresid student=sresid;
run;

proc gplot data=out2; symbol v=dot h=1;
     plot yresid*yhat/caxis=red ctext=blue vref=0;
title "population=year";

proc gplot data=out4; symbol v=dot h=1;
```

```
      plot yresid*yhat/caxis=red ctext=blue vref=0;
title "population=year+year*year";

proc gplot data=out6; symbol v=dot h=1;
      plot yresid*yhat/caxis=red ctext=blue vref=0;
title "population=year+year*year +Z";
run;
```

The above regression analysis of the US population over years can be performed using the free software R. One advantage of software R over SAS is that R generates regression diagnosis graphs relatively easily. We present the following R code that perform the regression analysis of the US population over years and generate all regression diagnosis plots in postscript format for different regression models.

```
year<-c(1790,1800,1810,1820,1830,1840,1850,1860,1870,1880,
        1890,1900,1910,1920,1930,1940,1950,1960,1970)
pop<-c(3.929,5.308,7.239,9.638,12.866,17.069,23.191,31.443,
       39.818,50.155, 62.947,75.994,91.972,105.710,122.775,
       131.669,151.325,179.323,203.211)
postscript("C:\\uspop.eps",horizontal=FALSE, onefile= FALSE,
           print.it=FALSE)

par(mfrow=c(2, 2))
plot(pop~year, pch=20, font=2, font.lab=2,
     ylab="population",xlab="Year",
     main="populations by Year")

fit<-lm(pop~year)
fitted<-fit$fitted
resid<-fit$residual
plot(fitted, resid, pch=20, cex=1.5, font=2, font.lab=2,
     ylab="Residual", xlab="Fitted Values",
     main="Population=Year")

yyear<-year*year
fit1<-lm(pop ~ year + yyear)
fitted1<-fit1$fitted
resid1<-fit1$residual
```

```
plot(fitted1, resid1, pch=20, cex=1.5, font=2, font.lab=2,
     ylab="Residual", xlab="Fitted Values",
     main="population=Year+Year^2")

z<-ifelse((year==1940)|(year==1950), 1, 0)
fit2<-lm(pop ~ year + yyear +z)
fitted2<-fit2$fitted
resid2<-fit2$residual
plot(fitted2, resid2, pch=20, cex=1.5, font=2, font.lab=2,
     ylab="Residual", xlab="Fitted Values",
     main="population=Year+Year^2+ Z")
dev.off()
```

3.20.3 *Studentized Residuals*

Without normalization the usual residual $e_i = y_i - \hat{y}_i$ is subject to the scale of the response y_i. It is inconvenient when several regression models are discussed together. We then consider the normalized residual. Since $\mathrm{Var}(e_i) = (1 - h_{ii})\sigma^2$ the normalized regression residual can be defined as

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}. \tag{3.71}$$

This normalized residual is called the studentized residual. Note that σ is unknown and it can be estimated by s. The studentized residual is scale-free and can be used for checking model assumption. Also it can be used for model diagnosis. If several regression models need to be compared the scale-free studentized residuals is a better measurement for model comparison.

3.20.4 *PRESS Residual*

The PRESS residual is the leave-one-out residual. To obtain the PRESS residual we fit the regression model without using the ith observation and calculate the fitted value from that model

$$\hat{y}_{i,-i} = \boldsymbol{x}_i b_{-i},$$

where b_{-i} is the least squares estimate of regression parameters without using the ith observation. $\hat{y}_{i,-i}$ is the fitted value calculated from the regression model without using the ith observation. The ith PRESS residual is defined as

$$e_{i,-i} = y_i - \hat{y}_{i,-i} = y_i - x_i b_{-i}. \tag{3.72}$$

The PRESS residual is the measurement of influential effect of the ith observation on the regression model. If the ith observation has a small influence on the regression model then $\hat{y}_i$ should be fairly close to $\hat{y}_{i,-i}$, therefore, the PRESS residual $e_{i,-i}$ should be close to the usual residual e_i. In order to discuss the PRESS residual and establish the relationship between usual the residual e_i and the PRESS residual $e_{i,-i}$ we first introduce the following the theorem (see Rao, 1973).

Theorem 3.16. *Let A be a nonsingular square $p \times p$ matrix and z be a p-dimensional column vector. The matrix $(A - zz^{'})^{-1}$ is given by*

$$(A - zz^{'})^{-1} = A^{-1} + \frac{A^{-1}zz^{'}A^{-1}}{1 - z^{'}A^{-1}z}. \tag{3.73}$$

The proof of the theorem is to directly show that $A - zz^{'}$ multiply the matrix on the right side of the above formula yields an identity matrix. This theorem will be used later to establish the relationship between the PRESS residual and the ordinary residual. For regression model $y = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$, write $\boldsymbol{X}$ as $(1, \boldsymbol{x}_2, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)$, where $\boldsymbol{x}_i$ is an n-dimensional vector. It is easy to verify that

$$\boldsymbol{X}^{'}\boldsymbol{X} = \begin{pmatrix} n & 1^{'}\boldsymbol{x}_1 & 1^{'}\boldsymbol{x}_2 & \cdots & 1^{'}\boldsymbol{x}_p \\ 1^{'}\boldsymbol{x}_1 & \boldsymbol{x}_1^{'}\boldsymbol{x}_1 & \boldsymbol{x}_1^{'}\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_1^{'}\boldsymbol{x}_p \\ 1^{'}\ \boldsymbol{x}_2 & \boldsymbol{x}_1^{'}\boldsymbol{x}_1 & \boldsymbol{x}_2^{'}\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_2^{'}\boldsymbol{x}_p \\ \cdots & & & & \\ 1^{'}\boldsymbol{x}_p & \boldsymbol{x}_2^{'}\boldsymbol{x}_p & \boldsymbol{x}_3^{'}\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_p^{'}\boldsymbol{x}_p \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_j x_{1j} & \sum_j x_{2i} & \cdots & \sum_j x_{pj} \\ \sum_j x_{1j} & \sum_j x_{1j}^2 & \sum_j x_{1j}x_{2j} & \cdots & \sum_j x_{1j}x_{pj} \\ \sum_j x_{2j} & \sum_j x_{2j}x_{2j} & \sum_j x_{2j}^2 & \cdots & \sum_j x_{2j}x_{pj} \\ \cdots & & & & \\ \sum_j x_{pj} & \sum_j x_{pj}x_{pj} & \sum_j x_{pj}x_{2j} & \cdots & \sum_j x_{pj}^2 \end{pmatrix}.$$

Remove the ith observation from $\boldsymbol{X}$ and perform the matrix multiplication of $\boldsymbol{X}^{'}_{-i}\boldsymbol{X}_{-i}$ we have

$$\boldsymbol{X}^{'}_{-i}\boldsymbol{X}_{-i} = \begin{pmatrix} n-1 & \sum_{j\neq i} x_{1j} & \sum_{j\neq i} x_{2j} & \cdots & \sum_{j\neq i} x_{pj} \\ \sum_{j\neq i} x_{1j} & \sum_{j\neq i} x_{1j}^2 & \sum_{j\neq i} x_{1j}x_{2j} & \cdots & \sum_{j\neq i} x_{1j}x_{pj} \\ \sum_{j\neq i} x_{2j} & \sum_{j\neq i} x_{2j}x_{2j} & \sum_{j\neq i} x_{2j}^2 & \cdots & \sum_{j\neq i} x_{2j}x_{pj} \\ \cdots & & & & \\ \sum_{j\neq i} x_{pj} & \sum_{j\neq i} x_{pj}x_{pj} & \sum_{j\neq i} x_{pj}x_{2j} & \cdots & \sum_{j\neq i} x_{pj}^2 \end{pmatrix}$$

$$= \boldsymbol{X}^{'}\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i^{'}.$$

Thus, we establish that

$$\boldsymbol{X}^{'}_{-i}\boldsymbol{X}_{-i} = \boldsymbol{X}^{'}\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i^{'}.$$

Using the formula above and set $A = \boldsymbol{X}^{'}\boldsymbol{X}$ we find

$$(\boldsymbol{X}^{'}_{-i}\boldsymbol{X}_{-i})^{-1} = (\boldsymbol{X}^{'}\boldsymbol{X} - \boldsymbol{x}_j\boldsymbol{x}_i^{'})^{-1}$$

$$= (\boldsymbol{X}^{'}\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}}{1 - \boldsymbol{x}_i^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{x}_i}$$

$$= (\boldsymbol{X}^{'}\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}}{1 - h_{ii}}$$

The following theorem gives the relationship between the PRESS residual and the usual residual.

Theorem 3.17. *Let regression model be $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. The relationship between the ith PRESS residual $e_{i,-i}$ and the ordinary ith residual e_i is given by*

$$e_{i,-i} = \frac{e_i}{1-h_{ii}} \tag{3.74}$$

$$Var(e_{i,-i}) = \frac{\sigma^2}{1-h_{ii}}. \tag{3.75}$$

Proof. For the regression model without using the ith observation the residual is

$$e_{i,-i} = \boldsymbol{y}_i - \boldsymbol{x}_i^{'} b_{-i} = \boldsymbol{y}_i - \boldsymbol{x}_i^{'}(\boldsymbol{X}_{-i}^{'}\boldsymbol{X}_{-i})^{-1}\boldsymbol{X}_{-i}^{'}\boldsymbol{y}_{-i}$$

$$\begin{aligned}
&= \boldsymbol{y}_i - \boldsymbol{x}_i^{'}\Big[(\boldsymbol{X}^{'}\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}}{1-h_{ii}}\Big]\boldsymbol{X}_{-i}^{'}\boldsymbol{y}_{-i}\\
&= \frac{(1-h_{ii})\boldsymbol{y}_i - (1-h_{ii})\boldsymbol{x}_i^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}_{-i}^{'}\boldsymbol{y}_{-i} - h_{ii}\boldsymbol{x}_i^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}_{-i}^{'}\boldsymbol{y}_{-i}}{1-h_{ii}}\\
&= \frac{(1-h_{ii})\boldsymbol{y}_i - \boldsymbol{x}_i^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}_{-i}^{'}\boldsymbol{y}_{-i}}{1-h_{ii}}
\end{aligned}$$

Note that $\boldsymbol{X}_{-i}^{'}\boldsymbol{y}_{-i} + \boldsymbol{x}_i\boldsymbol{y}_i = \boldsymbol{X}^{'}\boldsymbol{y}$ we have

$$\begin{aligned}
e_{i,-i} &= \frac{(1-h_{ii})\boldsymbol{y}_i - \boldsymbol{x}_i^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}(\boldsymbol{X}^{'}\boldsymbol{y} - \boldsymbol{x}_i\boldsymbol{y}_i)}{1-h_{ii}}\\
&= \frac{(1-h_{ii})\boldsymbol{y}_i - \boldsymbol{x}_i^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y} + \boldsymbol{x}_i^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{y}_i}{1-h_{ii}}\\
&= \frac{(1-h_{ii})\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i + h_{ii}\boldsymbol{y}_i}{1-h_{ii}} = \frac{\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i}{1-h_{ii}} = \frac{e_i}{1-h_{ii}}
\end{aligned}$$

For variance of PRESS residual $\text{Var}(e_{i,-i})$ we have

$$\text{Var}(e_{i,-i}) = \text{Var}(e_i)\frac{1}{(1-h_{ii})^2} = [\sigma^2(1-h_{ii})]\frac{1}{(1-h_{ii})^2} = \frac{\sigma^2}{1-h_{ii}} \qquad \square$$

The ith standardized PRESS residual is

$$\frac{e_{i,-i}}{\sigma_{i,-i}} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}. \tag{3.76}$$

3.20.5 *Identify Outlier Using PRESS Residual*

The standardized PRESS residual can be used to detect outliers since it is related to the ith observation and is scale free. If the ith PRESS residual is large enough then the ith observation may be considered as a potential outlier. In addition to looking at the magnitude of the ith PRESS residual, according to the relationship between the PRESS residual $e_{i,-i}$ and the regular residual e_i, the ith observation may be a potential outlier if the leverage h_{ii} is close to 1.

We now discuss how to deal with outlier in regression model. First, what is an outlier? An outlier is an observation at which the fitted value is not close enough to the observed response. i.e., there is breakdown in the model at the ith observation such that the location of the response is shifted. In this situation, the ith data point could be a potential outlier. To mathematically formulate this mean shift or model breakdown, we can write $E(\varepsilon_i) = \Delta \neq 0$. i.e., there is a non-zero mean shift in error term at the ith observation. If we believe that the choice and model assumptions are appropriate, it is suspectable that the ith observation might be an outlier in terms of the shift of the response from the model at that observation.

Another aspect of an outlier is that at the ith data point the $\text{Var}(\varepsilon)$ exceeds the error variance at other data points. i.e., there might be an inflation in variance at the ith observation. If the equal variance assumption is appropriate we may consider the ith observation as an outlier if the variance is inflated at the ith observation. So, outlier could be examined by checking both the mean response shift and the variance inflation at the ith data point. If equal variance assumption is no longer appropriate in the regression model we can use the generalized least squares estimate where the equal variance assumption is not required. The generalized least squares estimate will be discussed later.

A convenient test statistic used to detect outlier in regression model is the ith PRESS residual

$$e_{i,-i} = y_i - \hat{y}_{i,-i}.$$

If there is a mean shift at the ith data point, then we have

$$E(y_i - \hat{y}_{i,-i}) = E(e)i, -i = \Delta_i > 0.$$

Similarly, if there is a variance inflation at the ith data point we would like to use the standardized PRESS residual

$$\frac{e_{i,-i}}{\sigma_{i,-i}} = \frac{e_i}{\sigma\sqrt{1 - h_{ii}}}$$

to detect a possible outlier. Since σ is unknown, we can replace σ by its estimate s to calculate the standardized PRESS residual. Note that in the presence of a mean shift outlier s is not an ideal estimate of true standard deviation of σ. If we consider the situation where there is a mean shift outlier, the sample standard deviation s is biased upward, and is not an ideal estimate of standard error σ. One way to cope with it is to leave the

ith observation out and calculate the leave-one-out sum of squared residuals s_{-i}. It can be shown that the relationship between s_{-i} and regular s is

$$s_{-i} = \sqrt{\frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}}. \tag{3.77}$$

Replacing σ with s_{-i} we can construct a test statistic

$$t_i = \frac{y_i - \hat{y}_i}{s_{-i}\sqrt{1-h_{ii}}} \sim t_{n-p-1}. \tag{3.78}$$

Under the null hypothesis $H_0 : \Delta_i = 0$, the above test statistic has the centralized t distribution with degrees of freedom $n-p-1$, where n is the sample size and $p+1$ is the total number of parameters in the regression model. This test statistic can be used to test the hypothesis $H_0 : \Delta_i = 0$ versus the alternative $H_1 : \Delta_i \neq 0$. The above statistic is often called the R-student statistic. It tends larger if the ith data point is a mean shift outlier. Note that the two-tailed t-test should be used to test a mean shift outlier using the R-student statistic.

The R-student statistic can also be used to test variance inflation at the ith observation. If there is inflation in variance at the ith observation we should have $\text{Var}(\varepsilon_i) = \sigma^2 + \sigma_i^2$. Here σ_i^2 represents the increase in variance at the ith data point. The hypothesis may be defined as $H_0 : \sigma_i^2 = 0$ versus $H_1 : \sigma_i^2 \neq 0$. Note that the two-tailed t-test should be used as well.

3.20.6 *Test for Mean Shift Outlier*

Example 3.3. The coal-cleansing data will be used to illustrate the mean shift outlier detection in multiple regression. The data set has three independent variables. Variable x_1 is the percent solids in the input solution; x_2 is the pH value of the tank that holds the solution; and x_3 is the flow rate of the cleansing polymer in ml/minute. The response variable y is the measurement of experiment efficiency. The data set is presented in Table 3.11.

We first fit the coal-cleansing data to the multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

The SAS procedure REG is used to calculate the parameter estimates, HAT matrix, ordinary residuals, and R-student residuals for the above regression model. The program is presented as follows:

Table 3.11 Coal-cleansing Data

Experiment	x_1	x_2	x_3	y
1	1.5	6.0	1315	243
2	1.5	6.0	1315	261
3	1.5	9.0	1890	244
4	1.5	9.0	1890	285
5	2.0	7.5	1575	202
6	2.0	7.5	1575	180
7	2.0	7.5	1575	183
8	2.0	7.5	1575	207
9	2.5	9.0	1315	216
10	2.5	9.0	1315	160
11	2.5	6.0	1890	104
12	2.5	6.0	1890	110

```
proc reg data=coal outest=out1 tableout;
     model y=x1 x2 x3;
     output out=out2
     p=yhat r=resid h=hat rstudent=Rresid;
     run;
```

The fitted regression model is found to be

$$\hat{y} = 397.087 - 110.750x_1 + 15.5833x_2 - 0.058x_3.$$

The estimates of the regression parameters and the corresponding P-values are presented in Table 3.12.

Table 3.12 Parameter Estimates for Regression Model for Coal–Cleansing Data

Type	RMSE	Intercept	x1	x2	x3
PARMS	20.8773	397.087	-110.750	15.5833	-0.05829
STDERR	20.8773	62.757	14.762	4.9208	0.02563
T	20.8773	6.327	-7.502	3.1668	-2.27395
PVALUE	20.8773	0.000	0.000	0.0133	0.05257
L95B	20.8773	252.370	-144.792	4.2359	-0.11741
U95B	20.8773	541.805	-76.708	26.9308	0.00082

Before the analysis there was suspicion by the experimental engineer that the 9th data point was keyed in erroneously. We first fit the model without deleting the 9th data point. The fitted responses, residuals, values of diagonal elements in the HAT matrix, and values of the R-student statistic associated with each observation are calculated and listed in Table 3.13. The largest residual is the 9th residual ($e_9 = 32.192$) and the corresponding R-student statistic value is 2.86951, which implies that the 9th residual is greater than zero statistically. This finding would support the suspicion that the 9th data point was originally keyed in incorrectly.

Table 3.13 Residuals

Experiment	y_i	$\hat{y}_i$	e_i	h_{ii}	t_i
1	243	247.808	-4.8080	0.45013	-0.29228
2	261	247.808	13.1920	0.45013	0.83594
3	244	261.040	-17.0400	0.46603	-1.13724
4	285	261.040	23.9600	0.46603	1.76648
5	202	200.652	1.3480	0.08384	0.06312
6	180	200.652	-20.6520	0.08384	-1.03854
7	183	200.652	-17.6520	0.08384	-0.86981
8	207	200.652	6.3480	0.08384	0.29904
9*	216	183.808	32.1920*	0.45013	2.86951*
10	160	183.808	-23.8080	0.45013	-1.71405
11	104	103.540	0.4600	0.46603	0.02821
12	110	103.540	6.4600	0.46603	0.40062

Note that the statistical analysis only confirms that the 9th data point does not fit the proposed regression model well. Therefore, it may be a potential mean shift outlier. The decision on whether or not keeping this data point in the model has to be made jointly by regression model diagnosis, rechecking the experimental data, and consulting with the engineer who collected the data.

In the example above the mean shift outlier is tested individually. If there are multiple mean shift outliers, we can test these mean shift outliers simultaneously. To do so the threshold is calculated by the t distribution with degrees of freedom $n-p-1$ and test level α is chosen to be $0.025/m$, where n=total number of observations, $p+1$=number of regression parameters in the model, and m is the number of potential outliers that need to be tested. For small data set one may choose $m = n$. The Manpower data will be used to illustrate the simultaneous test for multiple mean shift outliers. The data were collected from 25 office sites by U.S. Navy. The purpose of the regression analysis is to determine the needs for the manpower in Bachelor Officers Quarters. The 7 independent variables and the response variable y in the data set are

x_1: Average daily occupancy
x_2: Monthly average numbers of check-ins
x_3: Weekly hours of service desk operation
x_4: Square feet of common use area
x_5: Number of building wings
x_6: Operational berthing capacity
x_7: Number of rooms
y : Monthly man-hours

The data set is presented in Table 3.14:

Table 3.14 Manpower Data

Site	x1	x2	x3	x4	x5	x6	x7	y
1	2.00	4.00	4	1.26	1	6	6	180.23
2	3.00	1.58	40	1.25	1	5	5	182.61
3	16.60	23.78	40	1.00	1	13	13	164.38
4	7.00	2.37	168	1.00	1	7	8	284.55
5	5.30	1.67	42.5	7.79	3	25	25	199.92
6	16.50	8.25	168	1.12	2	19	19	267.38
7	25.89	3.00	40	0	3	36	36	999.09
8	44.42	159.75	168	0.60	18	48	48	1103.24
9	39.63	50.86	40	27.37	10	77	77	944.21
10	31.92	40.08	168	5.52	6	47	47	931.84
11	97.33	255.08	168	19.00	6	165	130	2268.06
12	56.63	373.42	168	6.03	4	36	37	1489.50
13	96.67	206.67	168	17.86	14	120	120	1891.70
14	54.58	207.08	168	7.77	6	66	66	1387.82
15	113.88	981.00	168	24.48	6	166	179	3559.92
16	149.58	233.83	168	31.07	14	185	202	3115.29
17	134.32	145.82	168	25.99	12	192	192	2227.76
18	188.74	937.00	168	45.44	26	237	237	4804.24
19	110.24	410.00	168	20.05	12	115	115	2628.32
20	96.83	677.33	168	20.31	10	302	210	1880.84
21	102.33	288.83	168	21.01	14	131	131	3036.63
22	274.92	695.25	168	46.63	58	363	363	5539.98
23	811.08	714.33	168	22.76	17	242	242	3534.49
24	384.50	1473.66	168	7.36	24	540	453	8266.77
25	95.00	368.00	168	30.26	9	292	196	1845.89

Data Source: Procedure and Analyses for Staffing Standards: Data Regression Analysis handbook (San Diego, California: Navy Manpower and Material Analysis Center, 1979).

The SAS program for the simultaneous outlier detection is provided as follows. In this example we choose $m = n = 25$ since data set is not too large and we can test all observations simultaneously in the data set.

```
proc reg data=manpow outest=out1 tableout;
     model y=x1 x2 x3 x4 x5 x6 x7;
     output out=out2
     p=yhat r=e h=h RSTUDENT=t ;
run;

data out2; set out2;
cutoff=-quantile('T', 0.025/50, 25-8-1);
if abs(t)> cutoff then outlier="Yes";
else outlier="No";
run;
```

The output with information on multiple mean shift outliers is presented in Table 3.15. Note that in this example we tested all data points ($n = 25$) simultaneously. The cutoff for identifying multiple mean shift outlier is $\alpha/2n = 0.025/25$ quantile from the t distribution with degrees of freedom $n - 1 -$ number of parameters $= 25 - 1 - 8 = 16$. In the output, “No” indicates that the corresponding observation is not a mean shift outlier and “Yes” means a mean shift outlier.

In Table 3.15, we detect outlier using all data points as a whole. This approach is based on rather conservative Bonferroni inequality, i.e., set the critical value to be $t_{\alpha/2n,n-p-1}$, where n is the total number of observations to be tested and p is the total number of parameters in the regression model. We use this approach in situation where individual outlier detection and residual plot do not provide us enough information on model fitting. Detection of outlier as a whole may tell us that even individually there is no evidence to identify an outlier, but as compare to other residuals in the overall data set, one residual may be more extreme than other. The idea behind this approach is that when we fit data to a model we would expect the model can provide satisfactory fitted values for all data points as a whole.

In this example we set $\alpha = 0.05$, $n = 25$, and $p = 8$. The cutoff for the test statistic is

$$-t_{\alpha/2n,n-p-1} = -t_{0.05/50,16} = 3.686155.$$

Table 3.15 Simultaneous Outlier Detection

Obs	y_i	$\hat{y}_i$	e_i	h_{ii}	t_i	Outlier
1	180.23	209.98	-29.755	0.25729	-0.07360	No
2	182.61	213.80	-31.186	0.16088	-0.07257	No
3	164.38	360.49	-196.106	0.16141	-0.45944	No
4	284.55	360.11	-75.556	0.16311	-0.17621	No
5	199.92	380.70	-180.783	0.14748	-0.41961	No
6	267.38	510.37	-242.993	0.15890	-0.57043	No
7	999.09	685.17	313.923	0.18288	0.75320	No
8	1103.24	1279.30	-176.059	0.35909	-0.47199	No
9	944.21	815.47	128.744	0.28081	0.32464	No
10	931.84	891.85	39.994	0.12954	0.09139	No
11	2268.06	1632.14	635.923	0.12414	1.55370	No
12	1489.50	1305.18	184.323	0.20241	0.44258	No
13	1891.70	1973.42	-81.716	0.08020	-0.18179	No
14	1387.82	1397.79	-9.966	0.09691	-0.02235	No
15	3559.92	4225.13	-665.211	0.55760	-2.51918	No
16	3115.29	3134.90	-19.605	0.40235	-0.05406	No
17	2227.76	2698.74	-470.978	0.36824	-1.33105	No
18	4804.24	4385.78	418.462	0.44649	1.25660	No
19	2628.32	2190.33	437.994	0.08681	1.00741	No
20	1880.84	2750.91	-870.070	0.36629	-2.86571	No
21	3036.63	2210.13	826.496	0.07039	2.05385	No
22	5539.98	5863.87	-323.894	0.78537	-1.60568	No
23	3534.49	3694.77	-160.276	0.98846	-5.24234	Yes
24	8266.77	7853.50	413.265	0.87618	3.20934	No
25	1845.89	1710.86	135.029	0.54674	0.42994	No

We then compare the absolute value of each t_i with this cutoff to determine whether the corresponding observation is a possible outlier. Assuming that the regression model is correctly specified, the comparison between this cutoff and each observation in the data set alerts that the 23th observation might be an outlier.

The following example demonstrates the detection of multiple mean shift outliers. The magnitude of mean shift at different data point may be different. The technique for multiple outlier detection is to create variables which take value 1 at these suspicious data point and 0 elsewhere. We need to create as many such columns as the number of suspicious outliers if we believe there are different mean shifts at those data points. This way, we can take care of different magnitudes of mean shift for all possible outliers. If we think some outliers are of the same mean shift then for these outlier we should create a dummy variable that takes value 1 at these outliers and

0 elsewhere.

In the following example, the data points with a larger value of variable x_2 are suspicious and we would like to consider multiple data points 15, 18, 22, 23, and 24 as possible multiple outliers. We first create 5 dummy variables D_1, D_2, D_3, D_4, and D_5 that takes value 1 at observations 15, 18, 22, 23, and 24, and value 0 for all other observations in the data set. We then include these dummy variables in the regression model. The SAS program for detecting the mean shift outliers is provided below.

```
data manpow; set manpow;
if _n_=15 then D15=1; else D15=0;
if _n_=18 then D18=1; else D18=0;
if _n_=22 then D22=1; else D22=0;
if _n_=23 then D23=1; else D23=0;
if _n_=24 then D24=1; else D24=0;
run;

Proc reg data=manpow;
     model y=x1 x2 x3 x4 x5 x6 x7 D15 D18 D22 D23 D24;
run;
```

The output is presented in Table 3.16. The identified outlier is the 23th observation since the corresponding P-value is $0.0160 < 0.05$. Note that this time we identified the same outlier via multiple outlier detection approach.

Table 3.16 Detection of Multiple Mean Shift Outliers

Variable	df	b_i	std	t_i	$P > \|t\|$
Intercept	1	142.1511	176.0598	0.81	0.4351
x1	1	23.7437	8.8284	2.69	0.0197
x2	1	0.8531	0.8608	0.99	0.3412
x3	1	-0.2071	1.7234	-0.12	0.9063
x4	1	9.5700	16.0368	0.60	0.5618
x5	1	12.7627	21.7424	0.59	0.5681
x6	1	-0.2106	6.9024	-0.03	0.9762
x7	1	-6.0764	12.5728	-0.48	0.6376
Data15	1	723.5779	914.0654	0.79	0.4440
Data18	1	139.5209	611.3666	0.23	0.8233
Data22	1	-592.3845	900.1980	-0.66	0.5229
Data23*	1	-15354	5477.3308	-2.80	0.0160*
Data24	1	262.4439	1386.053	0.19	0.8530

3.21 Check for Normality of the Error Term in Multiple Regression

We now discuss how to check normality assumption on error term of a multiple regression model. It is known that the sum of squared residuals, divided by $n-p$, is a good estimate of the error variance, where n is the total number of observations and p is the number of parameters in the regression model, The residual vector in a multiple linear regression is given by

$$\boldsymbol{e} = (I - H)y = (I - H)(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (I - H)\boldsymbol{\varepsilon},$$

where H is the HAT matrix for this regression model. Each component $e_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j$. Therefore, the normality of residual is not simply the normality of the error term in the multiple regression model. Note that

$$\text{Cov}(\boldsymbol{e}) = (I - H)\sigma^2(I - H)' = (I - H)\sigma^2.$$

Hence we can write $\text{Var}(e_i) = (1 - h_{ii})\sigma^2$. If sample size is much larger than the number of the model parameters, i.e., $n >> p$, or sample size n is large enough, h_{ii} will be small as compared to 1, then $\text{Var}(e_i) \approx \sigma^2$. Thus, a residual in multiple regression model behaves like error if sample size is large. However, it is not true for small sample size. We point out that it is unreliable to check normality assumption using the residuals from a multiple regression model when sample size is small.

3.22 Example

In this section we provide some illustrative examples of multiple regression using SAS. The following SAS program is for calculating confidence intervals on regression mean and regression prediction. The Pine Tree data set used in this example is presented in Table 3.17.

The corresponding estimated of regression parameters for the model including all independent variables x_1, x_2, x_3 and the model including x_3, x_2 are presented in Tables 3.18 and 3.19. The confidence intervals on regression mean and regression prediction for the model including all variables and the model including x_1, x_2 are presented in Tables 3.20 and 3.21.

Table 3.17 Stand Characteristics of Pine Tree Data

Age	HD	N	MDBH
19	51.5	500	7.0
14	41.3	900	5.0
11	36.7	650	6.2
13	32.2	480	5.2
13	39.0	520	6.2
12	29.8	610	5.2
18	51.2	700	6.2
14	46.8	760	6.4
20	61.8	930	6.4
17	55.8	690	6.4
13	37.3	800	5.4
21	54.2	650	6.4
11	32.5	530	5.4
19	56.3	680	6.7
17	52.8	620	6.7
15	47.0	900	5.9
16	53.0	620	6.9
16	50.3	730	6.9
14	50.5	680	6.9
22	57.7	480	7.9

Data Source: Harold E, et al. "Yield of Old-field Loblolly Pine Plantations", Division of Forestry and Wildlife Resources Pub. FWS-3-72, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1972.

```
data pinetree; set pinetree;
x1= HD;
x2=age*N;
x3=HD/N;
run;

proc reg data=pinetree outest=out tableout;
model MDBH=x1 x2 x3/all;
run;

*Calculation of partial sum;
proc reg data=pinetree;
model MDBH=x1 x2 x3;
run;
```

Table 3.18 Parameter Estimates and Confidence Intervals Using x_1, x_2 and x_3

MODEL	TYPE	DEPVAR	RMSE	Intercept	x1	x2	x3
MODEL1	PARMS	MDBH	0.29359	3.23573	0.09741	-0.00017	3.4668
MODEL1	STDERR	MDBH	0.29359	0.34666	0.02540	0.00006	8.3738
MODEL1	T	MDBH	0.29359	9.33413	3.83521	-2.79003	0.4140
MODEL1	PVALUE	MDBH	0.29359	0.00000	0.00146	0.01311	0.6844
MODEL1	L95B	MDBH	0.29359	2.50085	0.04356	-0.00030	-14.2848
MODEL1	U95B	MDBH	0.29359	3.97061	0.15125	-0.00004	21.2185

Table 3.19 Parameter Estimates and Confidence Intervals after Deleting x_3

MODEL	TYPE	DEPVAR	RMSE	Intercept	x_1	x_2
MODEL1	PARMS	MDBH	0.28635	3.26051	0.1069	-0.00019
MODEL1	STDERR	MDBH	0.28635	0.33302	0.0106	0.00003
MODEL1	T	MDBH	0.28635	9.79063	10.1069	-5.82758
MODEL1	PVALUE	MDBH	0.28635	0.00000	0.0000	0.00002
MODEL1	L95B	MDBH	0.28635	2.55789	0.0846	-0.00026
MODEL1	U95B	MDBH	0.28635	3.96313	0.1292	-0.00012

```
proc reg data=pinetree outest=out tableout;
model MDBH=x1 x2;
run;

*Calculate fitted values and residuals;
proc reg data=pinetree;
model MDBH=x1 x2 x3;
output out=out  p=yhat  r=yresid  student=sresid
       LCLM=L_mean  UCLM=U_mean
       LCL=L_pred   UCL=U_pred;
run;

*Calculate fitted values and residuals after deleting X3;
proc reg data=pinetree;
model MDBH=x1 x2;
output out=out p=yhat r=yresid  student=sresid
       LCLM=L_mean  UCLM=U_mean
       LCL=L_pred   UCL=U_pred;
run;
```

If collinearity exists the regression analysis become unreliable. Although

Table 3.20 Confidence Intervals on Regression Mean and Prediction Without Deletion

Obs	MDBH	yhat	*Lmean*	*Umean*	*Lpred*	*Upred*	yresid	sresid
1	7.0	7.00509	6.69973	7.31046	6.31183	7.69835	-0.00509	-0.01991
2	5.0	5.29011	4.99935	5.58088	4.60316	5.97707	-0.29011	-1.11762
3	6.2	5.79896	5.53676	6.06115	5.12360	6.47431	0.40104	1.50618
4	5.2	5.55111	5.23783	5.86439	4.85433	6.24789	-0.35111	-1.38404
5	6.2	6.15311	5.92449	6.38173	5.49007	6.81615	0.04689	0.17171
6	5.2	5.07177	4.76288	5.38066	4.37695	5.76659	0.12823	0.50309
7	6.2	6.34891	6.17875	6.51908	5.70369	6.99414	-0.14891	-0.52731
8	6.4	6.21119	5.98863	6.43376	5.55021	6.87217	0.18881	0.68863

Table 3.21 Confidence Intervals on Regression Mean and Prediction After Deleting x_3

Obs	MDBH	yhat	*Lmean*	*Umean*	*Lpred*	*Upred*	yresid	sresid
1	7.0	6.96391	6.74953	7.17829	6.32286	7.60495	0.03609	0.13482
2	5.0	5.28516	5.00400	5.56632	4.61880	5.95151	-0.28516	-1.12512
3	6.2	5.82751	5.61623	6.03878	5.18749	6.46752	0.37249	1.38853
4	5.2	5.51907	5.26001	5.77813	4.86173	6.17641	-0.31907	-1.23345
5	6.2	6.14741	5.92731	6.36751	5.50443	6.79039	0.05259	0.19721
6	5.2	5.05755	4.76617	5.34892	4.38681	5.72828	0.14245	0.56791
7	6.2	6.34360	6.18055	6.50665	5.71785	6.96935	-0.14360	-0.52082
8	6.4	6.24510	6.10990	6.38030	5.62602	6.86418	0.15490	0.55504

we can identify highly dependent regressors and include one of them in the regression model to eliminate collinearity. In many applications, often it is rather difficulty to determine variable deletion. A simple way to combat collinearity is to fit the regression model using centralized data. The following example illustrates how to perform regression analysis on the centralized data using SAS. The regression model for centralized data is given by

$$y_i = \beta_0 + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \varepsilon_i.$$

We then create the centralize variables, $(x_{1i} - \bar{x}_1)$ and $(x_{2i} - \bar{x}_2)$, before performing the regression analysis. The following SAS code is for regression analysis using centralized data. The regression parameter estimators using the centralized data are presented in Table 3.22.

```
data example;
input yield temp time @@;
datalines;
77  180  1 79  160  2 82  165  1 83  165  2
```

```
85  170  1 88  170  2 90  175  1 93  175  2;
run;
*Centralize data;
proc means data=example noprint;
var temp time;
output out=aa mean=meantemp meantime;
run;

data aa; set aa;
call symput('mtemp', meantemp);
call symput('mtime', meantime);
run;

*Created centralized data ctime and ctemp;
data example; set example;
ctemp=temp-&mtemp;
ctime=time-&mtime;
run;

proc reg data=example outest=out1 tableout;
     model yield=ctemp ctime/noprint;
run;
```

Table 3.22 Regression Model for Centralized Data

Obs	MODEL	TYPE	DEPVAR	RMSE	Intercept	ctemp	ctime
1	MODEL1	PARMS	yield	5.75369	84.6250	0.37000	4.1000
2	MODEL1	STDERR	yield	5.75369	2.0342	0.36390	4.4568
3	MODEL1	T	yield	5.75369	41.6003	1.01678	0.9199
4	MODEL1	PVALUE	yield	5.75369	0.0000	0.35591	0.3998
5	MODEL1	L95B	yield	5.75369	79.3958	-0.56542	-7.3566
6	MODEL1	U95B	yield	5.75369	89.8542	1.30542	15.5566

The final multiple regression model is

$$\text{yield} = 84.625 + 0.37(\text{temperature} - 170) + 4.10(\text{time} - 1.5)$$

The test for linear hypothesis is useful in many applications. For linear regression models, SAS procedures GLM and MIXED are often used. The following SAS program uses the procedure GLM for testing linear hypothesis. Note that SAS procedure REG can also be used for testing linear hypothesis. The variable GROUP in the following example is a class variable. The results of the linear hypothesis tests are presented in Tables 3.23 and 3.24.

```
data example; input group weight HDL;
datalines;
 1 163.5 75.0
 ...
 1 144.0 63.5
 2 141.0 49.5
 ...
 2 216.5 74.0
 3 136.5 54.5
 ...
 3 139.0 68.0
 ;
 run;

*Regression analysis by group;
proc sort data=example;
by group;
run;

proc reg data=example outest=out1 tableout;
     model HDL=weight/noprint;
     by group;
run;

*Test for linear hypothesis of equal slopes;
proc glm data=example outstat=out1;
     class group;
     model HDL=group weight group*weight/ss3;
run;

proc print data=out1;
var _SOURCE_  _TYPE_  DF  SS  F PROB;
run;
```

We use the Pine Trees Data in Table 3.17 to illustrate how to test for

Table 3.23 Test for Equal Slope Among 3 Groups

SOURCE	TYPE	DF	SS	F	PROB
error	error	20	1712.36		
group	SS3	2	697.20	4.07157	0.03285
weight	SS3	1	244.12	2.85124	0.10684
weight*group	SS3	2	505.05	2.94946	0.07542

Table 3.24 Regression by Group

Group	Model	Type	Depvar	Rmse	Intercept	Weight
1	MODEL1	PARMS	HDL	7.2570	23.054	0.24956
1	MODEL1	STDERR	HDL	7.2570	25.312	0.15733
1	MODEL1	T	HDL	7.2570	0.911	1.58629
1	MODEL1	PVALUE	HDL	7.2570	0.398	0.16377
1	MODEL1	L95B	HDL	7.2570	-38.883	-0.13540
1	MODEL1	U95B	HDL	7.2570	84.991	0.63452
2	MODEL1	PARMS	HDL	10.3881	14.255	0.25094
2	MODEL1	STDERR	HDL	10.3881	17.486	0.11795
2	MODEL1	T	HDL	10.3881	0.815	2.12741
2	MODEL1	PVALUE	HDL	10.3881	0.446	0.07749
2	MODEL1	L95B	HDL	10.3881	-28.532	-0.03769
2	MODEL1	U95B	HDL	10.3881	57.042	0.53956
3	MODEL1	PARMS	HDL	9.6754	76.880	-0.08213
3	MODEL1	STDERR	HDL	9.6754	16.959	0.10514
3	MODEL1	T	HDL	9.6754	4.533	-0.78116
3	MODEL1	PVALUE	HDL	9.6754	0.002	0.45720
3	MODEL1	L95B	HDL	9.6754	37.773	-0.32458
3	MODEL1	U95B	HDL	9.6754	115.987	0.16032

linear hypothesis. The following SAS code test the linear hypothesis (a) H_0: $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$ versus H_1: at least one $\beta_i \neq \beta_j$. (b) H_0: $\beta_1 = \beta_2$ versus H_1: $\beta_1 \neq \beta_2$, both at a level 0.05 (default).

```
proc reg data=pinetree alpha=0.05;
model MDBH=x1 x2 x3;
test  intercept=0,
      x1=0,
      x2=0,
      x3=0;
test  x1=x2;
run;
```

The first hypothesis test (a) is the multiple test for checking if all parameters are zero and the observed value of the corresponding F test statistic is 2302.95 with the p-value $< .0001$. Thus, we cannot confirm H_0. For the second hypothesis test (b) the observed value of the corresponding F test statistic is 14.69 with the p-value 0.0015. Since the p-value is less than the significance level we cannot confirm H_0 either.
Again, we use the Pine Trees Data to illustrate how to find the least squares estimation under linear restrictions. The following SAS code compute the least squares estimates under the linear restrictions $\beta_0 = 3.23$ and $\beta_1 = \beta_2$.

```
proc reg data=pinetree;
model MDBH=x1 x2 x3;
restrict intercept=3.23, x1=x2/print;
run;
```

It is noted that the least squares estimates from the regression model without any linear restrictions are $b_0 = 3.2357$, $b_1 = 0.09741$, $b_2 = -0.000169$ and $b_3 = 3.4668$. The least squares estimates with the linear restrictions $\beta_0 = 3.23$ and $\beta_1 = \beta_2$ are $b_0 = 3.23$, $b_1 = b_2 = 0.00005527$ and $b_3 = 33.92506$.

Problems

1. Using the matrix form of the simple linear regression to show the unbiasness of the $\boldsymbol{b}$. Also, calculate the covariance of $\boldsymbol{b}$ using the matrix format of the simple linear regression.

2. Let $\boldsymbol{X}$ be a matrix of $n \times m$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$, where $\boldsymbol{X}_1$ is $n \times k$ matrix and $\boldsymbol{X}_2$ is $n \times (m-k)$ matrix. Show that

   (a). The matrices $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ are idempotent.

   (b). The matrix $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$ is idempotent.

   (c). Find the rank of the matrix $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$.

3. The least squares estimators of the regression model $Y = \boldsymbol{X\beta} + \varepsilon$ are linear function of the y-observations. When $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists the least squares estimators of $\boldsymbol{\beta}$ is $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}$. Let A be a constant matrix. Using $\text{Var}(Ay) = A\text{Var}(\boldsymbol{y})A'$ and $\text{Var}(\boldsymbol{y}) = \sigma^2 I$ to show that $\text{Var}(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

4. Show that the HAT matrix in linear regression model has the property $tr(H) = p$ where p is the total numbers of the model parameters.
5. Let h_{ii} be the ith diagonal elements of the HAT matrix. Prove that

   (a). For a multiple regression model with a constant term $h_{ii} \geq 1/n$.

   (b). Show that $h_{ii} \leq 1$. (Hint: Use the fact that the HAT matrix is idempotent.)

6. Assume that the data given in Table 3.25 satisfy the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,$$

where ε_i's are iid $N(0, \sigma^2)$.

Table 3.25 Data Set for Calculation of Confidence Interval on Regression Prediction

y	12.0	11.7	9.3	11.9	11.8	9.5	9.3	7.2	8.1	8.3	7.0	6.5	5.9
x_1	3	4	5	6	7	8	9	10	11	12	13	14	15
x_2	6	4	2	1	0	1	2	1	-1	0	-2	-1	-3

Data Source: Franklin A. Grabill, (1976), Theory and Application of the linear model. p. 326.

   (a). Find 80 percent, 90 percent, 95 percent, and 99 percent confidence interval for y_0, the mean of one future observation at $x_1 = 9.5$ and $x_2 = 2.5$.

   (b). Find a 90 percent confidence interval for $\bar{y}_0$, the mean of six observations at $x_1 = 9.5$ and $x_2 = 2.5$.

7. Consider the general linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \varepsilon$ and the least squares estimate $\boldsymbol{b} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}$. Show that

$$\boldsymbol{b} = \boldsymbol{\beta} + R\varepsilon,$$

where $R = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}$.

8. A scientist collects experimental data on the radius of a propellant grain (y) as a function of powder temperature, x_1, extrusion rate, x_2, and die temperature, x_3. The data is presented in Table 3.26.

   (a). Consider the linear regression model

$$y_i = \beta_0^{\star} + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \beta_3(x_{3i} - \bar{x}_3) + \epsilon_i.$$

   Write the vector $\boldsymbol{y}$, the matrix $\boldsymbol{X}$, and vector $\boldsymbol{\beta}$ in the model $\boldsymbol{y} = \boldsymbol{X\beta} + \varepsilon$.

Table 3.26 Propellant Grain Data

Grain Radius	Powder Temp (x_1)	Extrusion Rate (x_2)	Die Temp (x_3)
82	150	12	220
92	190	12	220
114	150	24	220
124	150	12	250
111	190	24	220
129	190	12	250
157	150	24	250
164	190	24	250

(b). Write out the normal equation $(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{y}$. Comment on what is special about the $\boldsymbol{X}'\boldsymbol{X}$ matrix. What characteristic in this experiment do you suppose to produce this special form of $\boldsymbol{X}'\boldsymbol{X}$.

(c). Estimate the coefficients in the multiple linear regression model.

(d). Test the hypothesis $H_0 : L\beta_1 = 0$, $H_0 : \beta_2 = 0$ and make conclusion.

(e). Compute $100(1-\alpha)\%$ confidence interval on $E(y|x)$ at each of the locations of x_1, x_2, and x_3 described by the data points.

(f). Compute the HAT diagonals at eight data points and comment.

(g). Compute the variance inflation factors of the coefficients b_1, b_2, and b_3. Do you have any explanations as to why these measures of damage due to collinearity give the results that they do?

9. For the data set given in Table 3.27

Table 3.27 Data Set for Testing Linear Hypothesis

y	x_1	x_2
3.9	1.5	2.2
7.5	2.7	4.5
4.4	1.8	2.8
8.7	3.9	4.4
9.6	5.5	4.3
19.5	10.7	8.4
29.3	14.6	14.6
12.2	4.9	8.5

(a). Find the linear regression model.

(b). Use the general linear hypothesis test to test

$$H_0 : \beta_1 = \beta_2 = 0$$

and make your conclusion. Use full and restricted model residual sums of squares.

10. Consider the general linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \varepsilon$ and the least squares estimate $\boldsymbol{b} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{y}$. Show that

$$\boldsymbol{b} = \boldsymbol{\beta} + R\varepsilon,$$

where $R = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}$.

11. In an experiment in the civil engineering department of Virginia Polytechnic Institute and State University in 1988, a growth of certain type of algae in water was observed as a function of time and dosage of copper added into the water. The collected data are shown in Table 3.28.

(a). Consider the following regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \varepsilon_i$$

Estimate the coefficients of the model, using multiple linear regression.

(b). Test $H_0 : \beta_{12} = 0$ versus $H_1 : \beta_{12} \neq 0$. Do you have any reason to change the model given in part (a).

(c). Show a partitioning of total degrees of freedom into those attributed to regression, pure error, and lack of fit.

(d). Using the model you adopted in part (b), make a test for lack of fit and draw conclusion.

(e). Plot residuals of your fitted model against x_1 and x_2 separately, and comment.

Table 3.28 Algae Data

y(unit of algae)	x_1(copper, mg)	x_2(days)
.3	1	5
.34	1	5
.2	2	5
.24	2	5
.24	2	5
.28	3	5
.2	3	5
.24	3	5
.02	4	5
.02	4	5
.06	4	5
0	5	5
0	5	5
0	5	5
.37	1	12
.36	1	12
.30	2	12
.31	2	12
.30	2	12
.30	3	12
.30	3	12
.30	3	12
.14	4	12
.14	4	12
.14	4	12
.14	5	12
.15	5	12
.15	5	12
.23	1	18
.23	1	18
.28	2	18
.27	2	18
.25	2	18

Table 3.28 Cont'd

y(unit of algae)	x_1(mg copper)	x_2(days)
.27	3	18
.25	3	18
.25	3	18
.06	4	18
.10	4	18
.10	4	18
.02	5	18
.02	5	18
.02	5	18
.36	1	25
.36	1	25
.24	2	25
.27	2	25
.31	2	25
.26	3	25
.26	3	25
.28	3	25
.14	4	25
.11	4	25
.11	4	25
.04	5	25
.07	5	25
.05	5	25