8.1 The General Linear Model and Least Squares. Parametric statistical inference deals in very large part with the general linear model, to which this chapter is devoted. In this section we shall lay out its definition, then we shall define the least squares problem for the unknown parameters of interest and present its solution, and we shall give a few introductory examples of the model. The definition is a bit on the wordy side, but it must be mastered before proceeding.

The general linear model can be written as

$$Y = X\beta + Z,$$

for which $Y$, $X$, $\beta$ and $Z$ satisfy the following:

1. $Y$ is an $n$-dimensional vector of observable random variables.
2. $X = (x_{ij})$ is an $n \times k$ matrix of known numbers, where $1 \leq k < n$, that satisfies rank$(X) = k < n$. $X$ is frequently called the design matrix of the linear model.
3. $\beta$ is a $k$-dimensional vector of unknown constants. The value of $\beta$ will be the central problem.
4. $Z$ is an $n$-dimensional vector of independent random variables whose coordinates $Z_1, \cdots, Z_n$ are unobservable, independent, identically distributed random variables whose common distribution is $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 > 0$ is unknown.

From this definition, we may determine that the $i$th coordinate, $Y_i$, of $Y$, can be expressed as

$$Y_i = \sum_{j=1}^{k} \beta_j x_{ij} + Z_i.$$

From the definition of the linear model given above, we see that $Y$ is $\mathcal{N}_n(X\beta, \sigma^2 I_n)$. Let us give two simple examples of linear models. The simplest occurs in what we shall refer to as the one sample problem of statistical inference. In this problem, one has $n > 1$ observable random variables $Y_1, \cdots, Y_n$ that are independent and identically distributed with common distribution $\mathcal{N}(\mu, \sigma^2)$, where both parameters $\mu$ and $\sigma^2 > 0$ are unknown.
If we denote \( 1_n \) as the \( n \)-dimensional vector of which all coordinates are 1’s, and if we let \( Z_i = Y_i - \mu \) for \( 1 \leq i \leq n \), then we easily see that \( Y = X\beta + Z \) where \( X = 1_n, \beta = (\mu) \) and \( k = 1 \).

Another example most frequently encountered is the two-sample problem. Consider observable random variables \( U_1, \ldots, U_s \) which are independent and identically distributed with common distribution \( N(\mu, \sigma^2) \), where \( \mu \) and \( \sigma^2 > 0 \) are unknown, and another set of observable random variables \( V_1, \ldots, V_t \) which are also independent and identically distributed with common distribution \( N(\nu, \sigma^2) \), where \( \nu \) is unknown, and further suppose that all \( s + t \) observable random variables are independent. Now we can express this as \( Y = X\beta + Z \), where \( Y = \begin{pmatrix} U \\ V \end{pmatrix} \). \( X \) is an \( (s + t) \times 2 \) matrix of which the first \( r \) rows are \((1, 0)\) and the last \( t \) rows are each \((0, 1)\), \( \beta = \begin{pmatrix} \mu \\ \nu \end{pmatrix} \) and \( Z_i = U_i - \mu \) for \( 1 \leq i \leq m \) and \( = V_i - \nu \) for \( s + 1 \leq i \leq s + t \). More complicated and useful linear models will appear in the next chapter.

We now turn our attention to the least squares problem. This problem is as follows: it is to find a vector function \( \tilde{\beta} = \tilde{\beta}(Y) \) of the observable random variables \( Y_1, \ldots, Y_n \) and the known constants in \( X \) such that \( ||Y - X \tilde{\beta}||^2 \) is minimized, i.e., if \( \tilde{\beta} \) is any other vector function of the observable random variables and the known constants in \( X \), then \( ||Y - X \tilde{\beta}||^2 \leq ||Y - X \beta||^2 \) for all \( \omega \in \Omega \), where \( ||x||^2 \) denotes the sum of the squares of the coordinates of \( x \). We shall need for this the following lemma.

**Lemma 1.** If \( A \) is an \( n \times k \) matrix with \( n \geq k \), and if \( \text{rank}(A) = k \), then \( \text{rank}(A^tA) = k \), i.e., \( A^tA \) is a nonsingular matrix.

**Proof:** If, to the contrary, \( \text{rank}(A^tA) < k \) then there exists a non-null vector \( z \in \mathbb{R}^k \) such that \( A^tAz = 0 \), and thus \( z^tA^tAz = 0 \). Hence \( ||Az||^2 = 0 \), which in turn implies \( Az = 0 \), i.e., \( \text{rank}(A) < k \), which contradicts the hypothesis.

**Theorem 1.** The value of \( \beta \) that minimizes \( ||Y - X \beta||^2 \) in the general linear model \( Y = X\beta + Z \) is \( \beta = (X^tX)^{-1}X^tY \).

**Proof:** We first note that by lemma 1, \( X^tX \) is nonsingular, and hence \( (X^tX)^{-1} \) exists. We must now show that the value of \( \beta \) that minimizes the value of \( ||Y - X \beta||^2 \) is \( \beta = (X^tX)^{-1}X^tY \). We may write

\[
||Y - X\beta||^2 = ||(Y - X\tilde{\beta}) + (X\tilde{\beta} - X\beta)||^2 = ||Y - X\tilde{\beta}||^2 + 2(Y - X\beta)^t(X\tilde{\beta} - X\beta) + ||X\tilde{\beta} - X\beta||^2.
\]

187
One half of the cross-product in this last expression is

\[ Y^tX\hat{\beta} - Y^tX\beta - \hat{\beta}^tX^tX\hat{\beta} + \hat{\beta}^tX^tX\beta \]

\[ = Y^tX\hat{\beta} - Y^tX\beta - Y^tX(X^tX)^{-1}X^tX(Y^tX)^{-1}X^tY + Y^tX(X^tX)^{-1}X^tX\beta \]

\[ = Y^tX(X^tX)^{-1}X^tY - Y^tX\beta - Y^tX(X^tX)^{-1}X^tY + Y^tX\beta = 0. \]

Thus \( ||Y - X\beta||^2 = ||Y - X\hat{\beta}||^2 + ||X(\hat{\beta} - \beta)||^2 \). It is easily seen that the unique value of \( \beta \) that minimizes the right hand side is \( \beta = \hat{\beta} \).

The unique solution \( \hat{\beta} = (X^tX)^{-1}X^tY \) of the least squares problem associated with the linear model is clearly an observable random vector, and one is tempted to use it as an estimate of \( \beta \). In the next theorem we show that it is an unbiased estimate of \( \beta \), i.e., \( E(\hat{\beta}) = \beta \), and we obtain its joint distribution.

**Theorem 2.** The joint distribution of \( \hat{\beta} \) is \( N_k(\beta, \sigma^2(X^tX)^{-1}) \) and thus is an unbiased estimate of \( \beta \).

**Proof:** As has been pointed out before, \( Y \) is \( N_n(X\beta, \sigma^2I_n) \). By theorem 5 in section 7.5, the distribution of \( \hat{\beta} = (X^tX)^{-1}X^tY \) is

\[ N_k((X^tX)^{-1}X^tX\beta, \sigma^2(X^tX)^{-1}X^tI_nX(X^tX)), \]

i.e., \( \hat{\beta} \) is \( N_n(\beta, \sigma^2(X^tX)^{-1}) \), which proves the theorem.

Following the definition of the linear model above, the two easiest examples of a linear model were exhibited. For these and others that will follow, if one wishes to estimate the unknown parameters \( \beta_1, \ldots, \beta_k \), theorems 1 and 2 indicate that upon observing the values of \( Y_1, \ldots, Y_n \), one should compute the least squares solution \( \hat{\beta} \) whose formula is given above.

**EXERCISES**

1. In the one sample problem given in this section, find the formula for the least squares solution of \( \mu \).
2. In the two sample problem given in this section, find the formulas for the least squares solutions for \( \mu \) and \( \nu \).
3. Very frequently in experimental work, a certain value of \( x \) is set by the experimenter, and he or she observes a response \( y \). In such a situation, the value of \( x \) is set deliberately, but the value of \( y \) that comes out is an observation on a random variable \( Y \). It is suspected many times that \( y \) is a
linear function of $x$, i.e., that $y = a + bx$, and thus that $Y = a + bx + Z$, where $Z$ is some error, possibly due to the instrument but also possibly due to other causes as well. So suppose that our experimenter sets $n > 2$ different values of $x$, call them $x_1, \cdots, x_n$, and for the $i$th value, he or she observes $Y_i$, where it is assumed that $Y_i = a + bx_i + Z_i$ and that $Z_1, \cdots, Z_n$ are independent and identically distributed with common distribution $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 > 0$ is unknown. What are $X$ and $\beta$ for this model? Also prove that

$$\hat{b} = \frac{\sum_{j=1}^{n} Y_j (x_i - \bar{x})}{\sum_{j=1}^{n} (x_i - \bar{x})^2}$$

and

$$\hat{a} = \bar{Y} - \hat{b}\bar{x},$$

where $\bar{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n)$ and $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$.

4. Prove without differentiating: if $x_1, \cdots, x_n$ are numbers and if $\bar{x}$ is defined by $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$, then the value of $x$ that minimizes $\sum_{j=1}^{n} (x_i - x)^2$ is $\bar{x}$.

### 8.2 The $F$-test of a Linear Hypothesis

Let us now consider the general linear hypothesis

$$Y = X\beta + Z,$$

where, as we defined it in the previous section, $Y$ is an $n$--dimensional vector of observable random variables, $X$ is an $n \times k$ matrix of known constants of rank $k < n$, $\beta$ is a $k$--dimensional vector of unknown constants (sometimes called parameters), and $Z$ is an $n$--dimensional random vector of unobservable random variables (sometimes called errors or white noise) that are independent, all with the same $\mathcal{N}(0, \sigma^2)$ distribution and where $\sigma^2 > 0$ is unknown. As we showed in the last section, $E(Y) = X\beta$. We first note that $E(Y)$ is a linear combination of the columns of $X$, i.e., if $X$ can be represented as the partitioned matrix of $k$ $n$--dimensional vectors $x_1, \cdots, x_k$, i.e.,

$$X = (x_1 \cdots x_k),$$

then $E(Y) = \sum_{j=1}^{k} \beta_j x_j$. Another way of saying this is as follows. Let $col.sp.(X)$ denote the set of all linear combinations of the columns of $X$. This forms a $k$--dimensional linear subspace of $\mathbb{R}^n$ with the columns of the matrix $X$ which are linearly independent, serving as a basis of $col.sp.(X)$.

The problem that we shall deal with is whether $E(Y)$ is actually in a certain predetermined proper linear subspace of $col.sp.(X)$. A basis of of this proper subspace would necessarily have fewer vectors than $k$, we might
denote the vectors of such a basis by \(x'_1, \ldots, x'_{k-q}\), where \(1 \leq q \leq k\). Thus if we denote \(X_0\) to be

\[X_0 = (x'_1 \cdots x'_{k-q}),\]

we may be concerned whether the correct model for the phenomenon that produced our observations might be

\[Y = X_0 \gamma + Z,\]

where \(Y\) and \(Z\) are as in the general linear model, \(\gamma\) is a \(k - q\) dimensional vector of unknown constants, and \(X_0\) is as defined above. We understand that \(\col.sp.(X_0) \subset \col.sp.(X)\), the set inclusion being proper. In brief, we are assuming that the general linear model is true, and we wish to test whether a more special linear model is also true.

Here is a simplest example of such a problem. In the two sample problem presented in the last section we most frequently wonder if \(\mu = \nu\). (We are not interested in what their common value is; we are just interested in whether the two parameters are equal.) In this case \(X_0\) is an \((s+t) \times 1\) matrix, \((1_{s+t})\) with \(\gamma = \gamma_1\), a one dimensional vector.

We now derive a test in the general case that the null hypothesis \(H_0: Y = X_0 \gamma + Z\) is the correct model against the alternative that \(H_0\) is not true but the general linear model is still true.

From a basic course in linear algebra, one knows that there exists an orthonormal basis \(\alpha_1, \ldots, \alpha_n\) of \(\mathbb{R}^n\) that satisfies the following: \(\alpha_1, \ldots, \alpha_{k-q}\) is an orthonormal basis of \(\col.sp.(X_0)\), and \(\alpha_1, \ldots, \alpha_k\) is an orthonormal basis of \(\col.sp.(X)\). Then, for every \(\omega \in \Omega\), there exist unique numbers \(W_1(\omega), \ldots, W_n(\omega)\) such that

\[Y(\omega) = \sum_{i=1}^{n} W_i(\omega) \alpha_i.\]

The numbers \(W_1(\omega), \ldots, W_n(\omega)\) are usually called the coordinates of the numerical vector \(Y(\omega)\) in the basis \(\alpha_1, \ldots, \alpha_n\) of \(\mathbb{R}^n\). Thus as functions over \(\Omega\), they are related by the equation

\[Y = (\alpha_1 \cdots \alpha_n)W,\]

where \(W\) is a vector function over \(\Omega\) whose \(i\)th coordinate is the function \(W_i\). It should be observed that the partitioned matrix \((\alpha_1 \cdots \alpha_n)\) is an orthogonal matrix, i.e.,

\[(\alpha_1 \cdots \alpha_n)^t(\alpha_1 \cdots \alpha_n) = (\alpha_1 \cdots \alpha_n)(\alpha_1 \cdots \alpha_n)^t = I_n.\]
Thus we may solve the equation $Y = (\alpha_1 \cdots \alpha_n)^t W$ for $W$ to obtain

$$W = (\alpha_1 \cdots \alpha_n)^t Y.$$ 

Since each $W_i$ is a linear combination of the random variables $Y_1, \cdots, Y_n$, it follows that each of the functions $W_1, \cdots, W_n$ is a random variable. In the traditional literature on this subject, these random variables are called the canonical variables.

**Lemma 1.** The following is true:

$$x_i^t \alpha_j = 0 \text{ for } 1 \leq i \leq k \text{ and } k + 1 \leq j \leq n,$$

and thus $X^t Y = \sum_{r=1}^k W_r X^t \alpha_r$.

*Proof:* Since, for $1 \leq i \leq k$, the column vector $x_i \in \text{col.sp.}(X)$, there exist constants $c_1, \cdots, c_k$ such that $x_i = \sum_{m=1}^k c_m \alpha_m$. Then for $k + 1 \leq j \leq n$, $x_i^t \alpha_j = \sum_{m=1}^k c_m \alpha_m^t \alpha_j = 0$ since $\alpha_1, \cdots, \alpha_n$ is an orthonormal basis of $\mathbb{R}^n$.

Thus

$$X^t Y = \sum_{r=1}^n W_r X^t \alpha_r = \sum_{r=1}^k W_r X^t \alpha_r.$$

**Lemma 2.** The random vector $X \hat{\beta}$ is the orthogonal projection of $Y$ onto $\text{col.sp.}(X)$, i.e., $X \hat{\beta} = \sum_{r=1}^k W_r \alpha_r$, and $Y - X \hat{\beta} = \sum_{r=k+1}^n W_r \alpha_r$.

*Proof:* We first note, since $\hat{\beta} = (X^t X)^{-1} X^t Y$, that

$$(X \hat{\beta})^t (Y - X \hat{\beta}) = \hat{\beta}^t X^t Y - \hat{\beta}^t X^t X \hat{\beta} = Y^t X (X^t X)^{-1} X^t Y - Y X (X^t X)^{-1} X^t X (X^t X)^{-1} X^t Y = 0.$$

Thus, $Y = X \hat{\beta} + (Y - X \hat{\beta})$ which is the sum of two $\omega-$pointwise orthogonal random vectors. The random vector $X \hat{\beta}$ is a linear combination of $\alpha_1, \cdots, \alpha_k$, and thus $Y - X \hat{\beta}$ must be a linear combination of $\alpha_{k+1}, \cdots, \alpha_n$. But $Y = \sum_{i=1}^n W_i \alpha_i$, and this representation is unique. Hence $X \hat{\beta} = \sum_{r=1}^k W_r \alpha_r$ and $Y - X \hat{\beta} = \sum_{r=k+1}^n W_r \alpha_r$.

**Theorem 1.** The distribution of $W$ is $\mathcal{N}_n(\varsigma, \sigma^2 I_n)$ where the last $n - k$ coordinates of $\varsigma$ are zeros, and thus $W_1, \cdots, W_n$ are independent.

*Proof:* Since $Y$ is $\mathcal{N}_n(X \beta, \sigma^2 I_n)$, and since $W = (\alpha_1 \cdots \alpha_n)^t Y$, it follows that $W$ is indeed multivariate normal with mean vector defined by $\varsigma = (\alpha_1 \cdots \alpha_n)^t X \beta$. The conclusion follows by applying lemma 1. Independence follows since $\text{Cov}(W) = \sigma^2 I_n$. 

191
Theorem 2. If \( s^2 \) is defined by
\[
s^2 = \frac{1}{n-k} \| \mathbf{Y} - \mathbf{X}\hat{\beta} \|^2 ,
\]
then \((n-k)s^2/\sigma^2\) has the chi-square distribution with \(n-k\) degrees of freedom, and
\[
\| \mathbf{Y} - \mathbf{X}\hat{\beta} \|^2 = \sum_{r=k+1}^{n} W_r^2 .
\]

Proof: By lemma 2, the second conclusion is immediate. By theorem 1, \( W_{k+1}, \ldots, W_n \) are independent, each being \( \mathcal{N}(0, \sigma^2) \). Thus \( \frac{W_{k+1}}{\sigma}, \ldots, \frac{W_n}{\sigma} \) are independent, and each is \( \mathcal{N}(0,1) \). By the definition of the chi-square distribution, we obtain \( (\frac{W_{k+1}}{\sigma})^2 + \cdots + (\frac{W_n}{\sigma})^2 \) has the chi square distribution with \(n-k\) degrees of freedom, which is equivalent to the first conclusion.

Theorem 3. The random vectors \( \hat{\beta} \) and \( \mathbf{Y} - \mathbf{X}\hat{\beta} \) are independent.

Proof: From lemma 2 above, we know that \( \mathbf{X}\hat{\beta} = \sum_{r=1}^{k} W_r \alpha_r \) and \( \mathbf{Y} - \mathbf{X}\hat{\beta} = \sum_{r=k+1}^{n} W_r \alpha_r \). Pre-multiplying both sides of the first of these equations by \( (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \), we obtain
\[
\hat{\beta} = \sum_{r=1}^{k} W_r (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \alpha_r .
\]

We observe that \( \hat{\beta} \) and \( \mathbf{Y} - \mathbf{X}\hat{\beta} \) are functions of disjoint subsets of the set \( W_1, \ldots, W_n \) of independent random variables, and thus by theorem 2 in section 7.3, they are independent.

Let us now return to the test of hypothesis that we laid out at the beginning of this section. Let us assume that not only is the general linear model true, but that the more specialized hypothesis \( H_0 : \mathbf{Y} = \mathbf{X}_0 \gamma + \mathbf{Z} \) is true. If this is the case, the least squares estimate of \( \gamma \) that minimizes \( \| \mathbf{Y} - \mathbf{X}_0 \gamma \|^2 \) is \( \hat{\gamma} = (\mathbf{X}'_0 \mathbf{X}_0)^{-1}\mathbf{X}'_0 \mathbf{Y} \), proved to be so in just the same way we proved that \( (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \mathbf{Y} \) is the least squares estimate of \( \beta \). And similarly, \( \mathbf{X}_0 \hat{\gamma} = \sum_{r=1}^{k-q} W_r \alpha_r \).

Theorem 4. If the null hypothesis \( H_0 : \mathbf{Y} = \mathbf{X}_0 \gamma + \mathbf{Z} \) is true, then the statistic \( \mathcal{F} \) defined by
\[
\mathcal{F} = \frac{\| \mathbf{X}\hat{\beta} - \mathbf{X}_0 \hat{\gamma} \|^2 n - k}{\| \mathbf{Y} - \mathbf{X}\hat{\beta} \|^2 q}
\]

192
has the \( F \)-distribution with \((q, n - k)\) degrees of freedom.

Proof: By what we have already shown, \( X\hat{\beta} - X_0\hat{\gamma} = \sum_{j=k-q+1}^{k} W_j \alpha_j \) and \( Y - X\hat{\beta} = \sum_{j=k+1}^{n} W_j \alpha_j \). Thus, by the independence of \( W_{k-q+1}, \ldots, W_n \), by theorem 2 in section 7.3, by theorem 3 above and by the fact that they are each \( \mathcal{N}(0, \sigma^2) \) if the null hypothesis is true, it follows that

\[
\frac{1}{\sigma^2} \| X\hat{\beta} - X_0\hat{\gamma} \|^2 = \sum_{j=k-q+1}^{k} \left( \frac{W_j}{\sigma} \right)^2
\]

and

\[
\frac{1}{\sigma^2} \| Y - X\hat{\beta} \|^2 = \sum_{j=k+1}^{n} \left( \frac{W_j}{\sigma} \right)^2
\]

are independent random variables whose distributions are chi-square with \( q \) and \( n - k \) degrees of freedom respectively. Thus by the definition of the \( F \)-distribution, the statistic

\[
\mathcal{F} = \frac{\| X\hat{\beta} - X_0\hat{\gamma} \|^2 n - k}{\| Y - X\hat{\beta} \|^2 q}
\]

has the \( F \)-distribution with \((q, n - k)\) degrees of freedom.

Theorem 4 is the big theorem, and it is used in tests of hypotheses as follows. The statistician observes a random vector \( Y \) whose mechanism is a linear model \( Y = X\beta + Z \). This is known or assumed to be the correct model for the observation. This model states, as we learned before, that the expectation vector, \( E(Y) \), is in the the linear subspace \( \text{col.sp.}(X) \). This much is assumed. One then wishes to test the null hypothesis that \( E(Y) \in \text{col.sp.}(X_0) \) which is a linear subspace of \( \text{col.sp.}(X_0) \). Suppose we observe the value of \( \mathcal{F} \) as defined above to be some positive number \( \mathcal{F}_0 \). We might ask ourselves: if our null hypothesis is true, what is the probability that \( \mathcal{F} \) can take a value as extreme as \( \mathcal{F}_0 \)? Note that when the null hypothesis is true, both \( X\hat{\beta} \) and \( X_0\hat{\gamma} \) are unbiased estimates of \( E(Y) \), so the norm of their difference should be small. So, for this problem the definition of the expression “as extreme as” should mean: what is the probability of observing a value of \( \mathcal{F} \) as large or larger than \( \mathcal{F}_0 \) when the null hypothesis \( E(Y) \in \text{col.sp.}(X_0) \) is true. This probability can be obtained from the density of the \( F_{q,n-k} \) distribution that we obtained in section 7.3. Such a probability can be found in standard tables published in the older texts or from readily available software.
EXERCISES

1. Prove that if \( U_1, \ldots, U_n \) are independent random variables, and if \( c_1, \ldots, c_n \) are non-zero constants then \( c_1 U_1, \ldots, c_n U_n \) are independent.

8.3 Statistical Inference on a Single Sample. By a single sample we mean \( n \) independent, identically distributed random variables, \( X_1, \ldots, X_n \), whose common distribution is \( \mathcal{N}(\mu, \sigma^2) \), where both \( \mu \) and \( \sigma^2 \) are unknown. At times questions are asked of a data set such as this, and some of these questions will be elaborated on in practical terms at the end of this section. Principal questions are whether the value of \( \mu \) is a known value \( \mu_0 \), what is a 100(1 - \( \alpha \))% confidence interval of \( \mu \), and whether a particular one of these observations, say, \( X_n \), is an outlier with respect to the rest of the sample. The results outlined below are concerned with the sample as defined above.

As we remarked earlier, this sample is easily verified to be a very special case of the linear hypothesis, 
\[
Y = X\beta + Z,
\]
where
\[
Y = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \quad X = 1_n, \quad \beta = (\mu), \quad \text{and} \quad Z = \begin{pmatrix} X_1 - \mu \\ \vdots \\ X_n - \mu \end{pmatrix}.
\]

One easily verifies that \( \hat{\beta} = \hat{\mu} = \bar{X} \), where \( \bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \) and \( \| Y - X\hat{\beta} \|^2 = (n - 1)s^2 \), where \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \).

**Theorem 1.** With the notation given above, the following are true,
\( \bar{X} \) is \( \mathcal{N}(\mu, \sigma^2/n) \).

*Proof:* This follows immediately from theorem 3 in section 7.2.

**Theorem 2.** With the notation given above, \( \bar{X} \) and \( s^2 \) are independent.

*Proof:* This follows immediately from theorem 3 in section 8.2.

**Theorem 3.** With the notation given above, \( (n - 1)s^2/\sigma^2 \) has the chi-square distribution with \( n - 1 \) degrees of freedom.

*Proof:* This follows from theorem 2 in section 8.2.

**Theorem 4.** The random variable
\[
T = \frac{\sqrt{n}(\bar{X} - \mu)}{s}
\]
has the \( t \)-distribution with \( n - 1 \) degrees of freedom.

*Proof:* By results already obtained, especially theorems 2 and 3 above, we know that \( \sqrt{n}(\bar{X} - \mu)/\sigma \) is \( \mathcal{N}(0, 1) \) and \( (n-1)s^2/\sigma^2 \) has the \( \chi^2_{n-1} \)-distribution.
We also know that both random variables are independent. Referring back to the definition of the $t$—distribution, we know that

$$\frac{\sqrt{n}(\overline{X} - \mu) / \sigma}{\sqrt{(n - 1)s^2/\sigma^2}/(n - 1)} = \frac{\sqrt{n}(\overline{X} - \mu)}{s}$$

has the $t_{n-1}$—distribution.

A problem of inference occasionally encountered is the desire to find a confidence interval of $\mu$ based on the observations $X_1, \cdots, X_n$, whose common distribution is $\mathcal{N}(\mu, \sigma^2)$, where both $\mu$ and $\sigma^2$ are unknown. This means that one desires to find two functions of $X_1, \cdots, X_n$, call them $L(X_1, \cdots, X_n)$ and $U(X_1, \cdots, X_n)$ that satisfy

$$P([L(X_1, \cdots, X_n) \leq \mu \leq U(X_1, \cdots, X_n)]) = 1 - \alpha,$$

where $\alpha$ is very small, like .01, .05 or even .10. So from the tables or software of the $t$—distribution, one finds the number $t_{1-\alpha/2}$ that satisfies $P([T \leq t_{1-\alpha/2}]) = 1 - \alpha/2$, where $T$ is a random variable that has the $t$—distribution with $n - 1$ degrees of freedom. We have proved in section 7.3 that that the $t$—distribution is symmetric, so

$$P \left( \frac{-t_{1-\alpha/2} \leq \sqrt{n}(\overline{X} - \mu)}{s} \leq t_{1-\alpha/2} \right) = 1 - \alpha.$$

This expression is easily shown to be equivalent to

$$P \left( \frac{\overline{X} - t_{1-\alpha/2}s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{1-\alpha/2}s \right) = 1 - \alpha.$$

Thus we may define

$$L(X_1, \cdots, X_n) = \overline{X} - \frac{t_{1-\alpha/2}s}{\sqrt{n}} \quad \text{and} \quad U(X_1, \cdots, X_n) = \overline{X} + \frac{t_{1-\alpha/2}s}{\sqrt{n}}.$$

This solves the problem stated above.

The problem of determining a confidence interval like that above arises in many situations. For example, a physicist might desire to determine the value of a certain physical constant. The constant has been measured $n$ times, each time producing a slightly different number due to random errors in the measuring instrument. The physicist will take $\overline{X}$ as the value of the
constant, but he or she will wish to know just how much error there is. The quantity $\frac{t_{1-a/2}}{\sqrt{n}}$ provides an answer to a certain extent.

Many times we might wish to test the null hypothesis that $\mu = \mu_0$, where $\mu_0$ is a known constant. This occurs in the problem of paired comparisons. An example of a paired comparison test is when two treatments, $T_1$ and $T_2$, tried on each of $n$ patients, with a desire to see which treatment if any gives a better (or larger) result. For patient number $i$, we try each treatment on the patient at different times; one such design is called the crossover trial. We may safely assume that the responses of each patient is independent from the responses of the others. Thus the responses $U_i$ and $V_i$ to treatments $T_1$ and $T_2$ respectively can be considered as a sequence of $n$ independent two-dimensional random vectors, $$\left( \begin{array}{c} U_1 \\ V_1 \\ \vdots \\ U_n \\ V_n \end{array} \right),$$ where each vector is assumed to be $N_2(\mu, C)$, where $\mu$ and $C$ are unknown. One might wish to test the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative that the two means are not equal. If we let $X_i = U_i - V_i$, $1 \leq i \leq n$, we have by theorem 2 in section 7.3 that $X_1, \ldots, X_n$ are independent and identically distributed with common distribution $N(\mu_1 - \mu_2, c_{11} + c_{22} - 2c_{12})$. Thus a test for $H_0 : \mu_1 = \mu_2$ is equivalent to testing that the common expectation of $X_1, \ldots, X_n$ is zero.

So suppose that we have our sample $X_1, \ldots, X_n$ of independent, identically distributed random variables whose common distribution is $N(\mu, \sigma^2)$, and we wish to test the null hypothesis $H_0 : \mu = \mu_0$, where the number $\mu_0$ is known. By theorem 4, we know that when $H_0$ is true, then

$$T = \frac{\sqrt{n}(X - \mu_0)}{s}$$

has the $t$–distribution with $n - 1$ degrees of freedom and is symmetric about zero. So suppose we observe that the value of $T$ is $t_0$. If $t_0 > 0$, we determine the probability that a random variable $W$ with the $t_{n-1}$ distribution is greater than $t_0$. If this probability is unbelievably small, say, less than .05 or less than .01, then we would reject the null hypothesis and declare that $\mu > 0$. If $t_0 < 0$, and if $P(|W \leq t_0|)$ is unbelievably small, we would reject the null hypothesis that $\mu = 0$ in favor of the alternative $\mu < 0$.

Another example of a one-sample test is if one has a great deal of historical data on the response to treatment $T_1$ and wishes to test the response to a new
treatment. So $n$ patients are selected, and their responses are observed, being assumed to come from some $\mathcal{N}(\mu, \sigma^2)$ distribution. The historical average, $\mu_0$, is known, and one wishes to know if $\mu = \mu_0$ or if $\mu > \mu_0$ or if $\mu < \mu_0$. The test outlined above provides us with a method of finding an answer.

We should also remark that the sign test presented in section 6.2 provides an alternative to the paired comparison $t$-test, although for reasons that we shall not go into here, if the data come from a normal distribution, the paired comparison $t$-test is better.

**EXERCISES**

1. Prove that in the linear model representation for the single sample, $\hat{\mu} = \overline{X}$ and where $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ and $\| Y - X\beta \|^2 = (n-1)s^2$, where

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

2. Prove that with the definition given for $t_{1-a/2}$ that

$$P \left( \left[ -t_{1-a/2} \leq \frac{\sqrt{n}(X - \mu)}{s} \leq t_{1-a/2} \right] \right) = 1 - \alpha.$$

3. Consider a two-dimensional random vector

$$\begin{pmatrix} U \\ V \end{pmatrix}$$

which is assumed to be $\mathcal{N}(\mu, C)$, and let $X = U - V$. Prove that $X$ is $\mathcal{N}(\mu_1 - \mu_2, c_{11} + c_{22} - 2c_{12})$.

**8.4 Statistical Inference on Two Samples.** Perhaps the most frequently used test in statistical inference is the two-sample $t$-test. It basically arises when two treatments are to be compared on two different groups of subjects. The problem is whether the two treatments offer the same results or whether one is better than another in a numerical way. We shall render the idea more precise by presenting its mathematical formulation.

Let us suppose that there are $m+n$ observable random variables denoted by $X_1, \cdots, X_m, Y_1, \cdots, Y_n$, all assumed to be independent. The $m$ random
variables \(X_1, \ldots, X_m\) are assumed to be independent and identically distributed with common distribution \(\mathcal{N}(\mu, \sigma^2)\). The random variables \(Y_1, \ldots, Y_n\) are assumed to be independent and identically distributed with common distribution \(\mathcal{N}(\nu, \tau^2)\). We wish to test the null hypothesis \(H_0: \mu = \nu\) against the alternative that this hypothesis is not true. This can be done easily if we are able to assume that \(\sigma^2 = \tau^2\), and in many cases this is possible.

So let us assume that the common distribution of \(Y_1, \ldots, Y_n\) is \(\mathcal{N}(\mu, \sigma^2)\). In other words, the variances of all \(m+n\) random variables are equal, but the common means or expectations \(\mu\) and \(\nu\) of the two samples might or might not be equal. We wish to test the null hypothesis \(H_0: \mu = \nu\) that the two means are equal. Note that we may express this as

\[
\left( \begin{array}{c} X \\ Y \\ \end{array} \right) = \left( \begin{array}{cc} 1_m & 0 \\ 0 & 1_n \end{array} \right) \left( \begin{array}{c} \mu \\ \nu \end{array} \right) + \left( \begin{array}{c} X - \mu 1_m \\ Y - \nu 1_n \end{array} \right)
\]

which satisfies the definition of the linear model where \(n\) is replaced by \(m+n\) and where the design matrix is clearly an \((m+n) \times 2\) matrix of rank 2. The least squares estimate of the two unknown parameters becomes

\[
\left( \begin{array}{c} \hat{\mu} \\ \hat{\nu} \end{array} \right) = \left( \left( \begin{array}{cc} 1_m & 0 \\ 0 & 1_n \end{array} \right)^t \left( \begin{array}{cc} 1_m & 0 \\ 0 & 1_n \end{array} \right) \right)^{-1} \left( \begin{array}{cc} 1_m & 0 \\ 0 & 1_n \end{array} \right)^t \left( \begin{array}{c} X \\ Y \end{array} \right) = \left( \begin{array}{c} \bar{X} \\ \bar{Y} \end{array} \right).
\]

Also, one easily shows that

\[
\left\| \left( \begin{array}{c} X \\ Y \end{array} \right) - \left( \begin{array}{cc} 1_m & 0 \\ 0 & 1_n \end{array} \right) \left( \begin{array}{c} \bar{X} \\ \bar{Y} \end{array} \right) \right\|^2 = (m-1)s_x^2 + (n-1)s_y^2,
\]

where \(s_x^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})^2\) and \(s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2\). By theorem 2 in section 8.3 and by theorem 2 in section 7.3, it follows that \(\bar{X}, \bar{Y}, s_x^2\) and \(s_y^2\) are independent random variables. Again, invoking theorem 2 of section 7.3 and results that we have proved about the normal distribution, we may conclude that

\[
\sqrt{\frac{mn}{m+n}((X - \mu) - (Y - \nu))/\sigma} \text{ and } \frac{1}{\sigma^2((m-1)s_x^2 + (n-1)s_y^2)}
\]

are independent random variables whose distributions are \(\mathcal{N}(0, 1)\) and \(\chi^2_{m+n-2}\) respectively. Thus, by the definition of the \(t\)-distribution in section 7.3, the random variable

\[
T = \sqrt{\frac{mn}{m+n} \frac{(X - \mu) - (Y - \nu)}{\sqrt{\frac{((m-1)s_x^2 + (n-1)s_y^2)}{m+n-2}}}}
\]

198
has the $t$-distribution with $m + n - 2$ degrees of freedom. This is not an observable random variable. However, suppose the null hypothesis $H_0 : \mu = \nu$ is true. Then it is observable, and

$$T = \sqrt{\frac{mn}{m+n}} \frac{\bar{X} - \bar{Y}}{\sqrt{(m-1)s_x^2 + (n-1)s_y^2}}$$

has the $t$-distribution with $m + n - 2$ degrees of freedom. We observe the value of $T$ from our data. If $T = t_0$, we might ask ourselves: what is the probability that a random variable $W$ with the $t_{m+n-2}$-distribution will be as extreme as this? If $t_0 > 0$, and if this probability is unbelievably small, we reject the null hypothesis and declare that $\mu > \nu$. If $t_0 < 0$, and if the probability is unbelievably small, then we reject the null hypothesis and declare that $\mu < \nu$. This test is called the $t$-test or the two-sample $t$-test.

The two sample $t$-test is valid only if the two variances are equal, i.e., $\sigma^2 = \tau^2$. How can one tell if $\sigma^2 = \tau^2$, when we don’t know what their values are? In truth, all we are interested in is whether they are equal, whatever their values might be. We might wish to test the null hypothesis $H_0 : \sigma^2 = \tau^2$ against the alternative hypothesis that these variances are not equal. Recall that $(m-1)s_x^2/\sigma^2$ and $(n-1)s_y^2/\tau^2$ were shown to be independent, both with chi-square distributions with $m - 1$ and $n - 1$ degrees of freedom respectively. Thus, when the null hypothesis $H_0 : \sigma^2 = \tau^2$ is true, it follows that the test statistic $V$ defined by

$$V = \frac{s_x^2}{s_y^2}$$

has the $F_{m-1,n-1}$ distribution, i.e., the $F$-distribution with $(m - 1, n - 1)$ degrees of freedom. In practice, we compute the value of $V$ from the data, call it $v_0$, and ask ourselves: if $H_0 : \sigma^2 = \tau^2$ is true, what is the probability that a random variable $S$ that has the $F_{m-1,n-1}$ distribution can have a value as extreme as $v_0$? If $P([S \geq v_0])$ is unbelievably small, then the null hypothesis should be rejected, and one should conclude that $\sigma^2 > \tau^2$. If $P([S \leq v_0])$ is unbelievably small, then the null hypothesis should be rejected, and one should conclude that $\sigma^2 < \tau^2$. If we do not reject the null hypothesis, then we may apply the two-sample $t$-test that is developed above. It should also be remarked that this test for equality of variances is applicable to other problems. Sometimes it is desired to determine if two measuring instruments not only give the same expected measurement of what it is used to measure
but also has the same accuracy. By accuracy, the experimenter obviously means that they have the same variance.

There are other tests for the null hypothesis $H_0 : \sigma^2 = \tau^2$. One of them has already been presented. It is the permutation test presented in section 6.5. It should be noted that the permutation test does not really depend on the two samples coming from normal populations. In that test, the null hypothesis is that the two samples have the same common distributions which includes the problem presented here.

EXERCISES

1. Verify that

\[
\left\| \begin{pmatrix} X \\ Y \end{pmatrix} - \begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \right\|^2 = (m - 1)s_x^2 + (n - 1)s_y^2,
\]

where $s_x^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})^2$ and $s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$.

8.5 One-Way Analysis of Variance. In section 8.4, the problem was to determine whether two samples on normal distributions with the same variances have the same expectations. We now extend the problem from two samples to any number, $k$, of samples. In what follows, when $k = 2$, the test turns out to be the same test as the two-sample $t$-test.

We consider here $k$ samples. The $i$th of these samples consists of $n_{i}$ observable random variables, $X_{i1}, \cdots, X_{in_{i}}$, which are assumed to be independent and identically distributed with common distribution $N(\mu_{i}, \sigma^2)$, $1 \leq i \leq k$. The $k+1$ parameters $\mu_1, \cdots, \mu_k, \sigma^2$ are all unknown. It is important to note that the variances of all the random variables are the same. We shall let $n = n_1 + \cdots + n_k$, and we shall assume that all $n$ random variables in $\{X_{ij}, 1 \leq j \leq n_{i}, 1 \leq i \leq k\}$ are independent. Our problem here is to test the null hypothesis

\[ H_0 : \mu_1 = \cdots = \mu_k \]
against the alternative that not all the $\mu_i$'s are equal. Let us denote

$$Y = \begin{pmatrix} X_{11} \\ \vdots \\ X_{1n_1} \\ X_{21} \\ \vdots \\ X_{2n_2} \\ \vdots \\ X_{k1} \\ \vdots \\ X_{kn_k} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}.$$ 

Note that the first row of the design matrix $X$ is repeated $n_1$ times, the $(n_1+1)$th row is repeated $n_2$ times, etc. It is clear that we may write

$$Y = X\beta + Z,$$

where $Z_1, \ldots, Z_n$ are independent $\mathcal{N}(0, \sigma^2)$ random variables. The null hypothesis becomes $H_0 : Y = X_0\gamma + Z$, where $X_0 = 1_n$ and $\gamma = (\mu)$ (a one-dimensional parameter vector). Also, from our general linear model, it is clear that $q = k - 1$ (the number of columns in $X$ minus the number of columns in $X_0$). Thus we observe the value $F_0$ of

$$F = \frac{\| X\widehat{\beta} - X_0\widehat{\gamma} \|^2 n - k}{\| Y - X\widehat{\beta} \|^2 k - 1}.$$ 

If this value is unbelievably large, we shall reject the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ in favor of the alternative that not all the $\mu_i$'s are equal.

The above formula for $F$ can also be written as

$$F = \frac{\frac{1}{k-1} \sum_{i=1}^k n_i (X_j - \overline{X})^2}{\frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \overline{X_i})^2},$$

where

$$\overline{X_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \quad \text{and} \quad \overline{X} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}.$$
This is the preferred formula for computing the value of $F$.

It should be remarked that the technique of setting up a permutation test that was used in section 5.3 is certainly applicable in this situation. If the null hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ is true, then the ordered set of random variables

$$X_{11}, \cdots, X_{1n_1}, \cdots, X_{k1}, \cdots, X_{kn_k}$$

are all independent and identically distributed $\mathcal{N}(\mu_1, \sigma^2)$. So upon observing the values

$$x_{11}, \cdots, x_{1n_1}, \cdots, x_{k1}, \cdots, x_{kn_k},$$

one computes the value of $F$ above, call it $F_0$, and then wonders: if the null hypothesis is true, then all rearrangements of these numbers are equally likely. The problem that the permutation test places before us is this. What proportion of these $n!$ permutations would yield a value as extreme, i.e., as large as $F_0$? In most situations, the value of $n$ and therefore of $n!$ is too large to compute. But one can make use of Bernoulli’s theorem to approximate this proportion or probability by taking a large number of random permutations of the ordered sequence of values displayed above, computing the value of $F$ for each permutation and finding the relative frequency of times in which the computed value of $F$ is as large or larger than $F_0$. It is recommended in practical work that both of these procedures be used.

There is yet another test for $H_0 : \mu_1 = \cdots = \mu_k$. It is called the Kruskal-Wallis test, and it will be developed in the chapter on nonparametric inference.

A further question arises that did not arise in the hypothesis testing that we did before. Suppose that the $P-$value of the $F-$test is unbelievably small, and thus the hypothesis $H_0 : \mu_1 = \cdots = \mu_k$ that all the population means are equal is rejected. Note that this does not mean that they are all distinct. It means that we have good reason to believe that for at least one pair of samples, $\mu_i \neq \mu_j$ for some $i \neq j$. In practical work, one is interested in more than which pairs are unequal; one is interested in ordering their values from smallest to largest. For example, if $k = 6$, it might be that in reality $\mu_2 = \mu_5 < \mu_6 = \mu_4 = \mu_1 < \mu_3$. This problem of ordering the values of these $\mu_i's$ after deciding to reject the null hypothesis is a complicated one and is part of the general area of statistical theory usually referred to as multiple comparisons. We shall treat it here rather briefly (and somewhat inadequately) by the method called Fisher’s least significant difference method. Notice that in our treatment above, $\hat{\beta}_i = \bar{X}_i$, and $\hat{\beta}_j = \bar{X}_j$, where $i \neq j$, are
independent and $\mathcal{N}(\mu_i, \sigma^2/n_i)$ and $\mathcal{N}(\mu_j, \sigma^2/n_j)$ respectively; they are also independent of

$$\frac{1}{\sigma^2} \| \mathbf{Y} - X\hat{\boldsymbol{\beta}} \|^2 = \frac{1}{\sigma^2} \sum_{r=1}^{k} (n_r - 1) s_r^2,$$

whose distribution is $\chi^2_{n-k}$, where $s_r^2 = \frac{1}{n_r} \sum_{u=1}^{n_r} (X_{ru} - \overline{X}_r)^2$. Thus

$$\sqrt{\frac{n_i n_j}{n_i + n_j}} \frac{(\overline{X}_i - \overline{X}_j) - (\mu_i - \mu_j)}{\sqrt{\frac{\sum_{r=1}^{k} (n_r - 1) s_r^2}{n_k - 1}}}$$

has the $t_{n-k}$-distribution. So we observe the values of all of these $\binom{k}{2}$ statistics, assuming that $\mu_i - \mu_j = 0$, i.e., we observe all $\binom{k}{2}$ values of

$$\sqrt{\frac{n_i n_j}{n_i + n_j}} \frac{\overline{X}_i - \overline{X}_j}{\sqrt{\frac{\sum_{r=1}^{k} (n_r - 1) s_r^2}{n_k - 1}}}$$

If the value of any particular one of these is unbelievably large (with respect to the $t_{n-k}$-distribution), it is reasonable for us to say that $\mu_i > \mu_j$. If the value of any particular one of these is unbelievably small (with respect to the $t_{n-k}$-distribution), it is reasonable for us to say that $\mu_i < \mu_j$. Warning: one does not always get a consistent set of pairwise equalities and inequalities, and sometimes one must adjust what is believable and what is unbelievable to obtain a consistent such set. We shall leave this problem go at that.

**EXERCISES**

1. Verify the assertion made in this section that

$$\frac{\| X\hat{\boldsymbol{\beta}} - X_0 \hat{\boldsymbol{\gamma}} \|^2}{\| \mathbf{Y} - X\hat{\boldsymbol{\beta}} \|^2} = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\overline{X}_j - \overline{X})^2$$

2. Prove that $\text{rank}(X) = k$ and $\text{rank}(X_0) = 1$.

3. Prove that

$$\sqrt{\frac{n_i n_j}{n_i + n_j}} \frac{(\overline{X}_i - \overline{X}_j) - (\mu_i - \mu_j)}{\sqrt{\frac{\sum_{r=1}^{k} (n_r - 1) s_r^2}{n_k - 1}}}$$

203
has the \( t \)-distribution with \( n - k \) degrees of freedom.

### 8.6 Two-Factor Designs with No Interaction

In section 8.5, we considered the case of \( k \) independent samples, all with the same population variances but with possibly different population means, and we addressed the problem of whether the population means were equal or not. We might have thought of these different populations as really the same population but with possibly different levels of some treatment. Our problem then became that of determining if the responses were essentially different for the different levels of treatment or that the responses did not depend on the different levels. In a two-factor design with no interaction, we do not consider just one factor as before but two factors. The second factor might be a subdivision of the population into different groups or strata. The simplest model of this sort is where one feels that it is safe to assume that a difference between two levels of the first factor is the same for all levels of the second factor. The simplest model of this is what is called the two-factor design with no interaction and with one observation per cell. That is, there is one subject at level \( i \) of the first factor and at level \( j \) of the second factor, where \( 1 \leq i \leq r, 1 \leq j \leq c \), where \( r \) denotes the number of levels of the first factor and \( c \) denotes the number of levels of the second factor. The response from this subject will be denoted by \( X_{ij} \), an observable random variable, and the model considered is that

\[
X_{ij} = a_i + b_j + Z_{ij},
\]

where \( \{Z_{ij} : 1 \leq i \leq r, 1 \leq j \leq c\} \) are independent, unobservable random variables that are independent and identically distributed with common distribution \( \mathcal{N}(0, \sigma^2) \), where \( \sigma^2 > 0 \) is unknown. There is one other restriction that we must place on this. Note that if \( a_i \) were replaced by \( a'_i = a_i + \Lambda \) and if \( b_i \) were replaced by \( b'_i = b_i - \Lambda \), then the equation \( E(X_{ij}) = a_i + b_j = a'_i + b'_j \) holds, thus denying us a unique representation of \( E(X_{ij}) \) in terms of the parameters. However, if we are able to place a restriction that the \( b_j \)'s that they must satisfy \( b_1 + \cdots + b_c = 0 \), then from among all such representations of the \( E(X_{ij})'s \), we should have a unique representation of \( E(X_{ij}) = a_i + b_j \).

We may sum this up and prove it with the following theorem.

**Theorem 1.** If \( M_{ij} = a'_i + b'_j \) for \( 1 \leq i \leq r, 1 \leq j \leq c \), then there exists exactly one set of numbers \( \{a_1, \ldots, a_r, b_1, \ldots, b_c\} \) such that \( M_{ij} = a_i + b_j \) for \( 1 \leq i \leq r, 1 \leq j \leq c \) and such that \( b_1 + \cdots + b_c = 0 \). Furthermore, for all \( 1 \leq i < j \leq r \) and all \( 1 \leq u < v \leq c \), \( a_i - a_j = a'_i - a'_j \) and \( b_u - b_v = b'_u - b'_v \).
Proof: For every $1 \leq i \leq r, 1 \leq j \leq c$, let us define $a_i = a_i' + \bar{b}$ and $b_j = b_j' - \bar{b}$, where $\bar{b} = (b_1' + \cdots + b_c')/c$. One can easily verify that \{$a_1, \ldots, a_r, b_1, \ldots, b_c$\} so defined satisfy the requirement that $M_{ij} = a_i + b_j$ for $1 \leq i \leq r, 1 \leq j \leq c$ and $b_1 + \cdots + b_c = 0$ and such that for all $1 \leq i < j \leq r$ and all $1 \leq u < v \leq c$, the following equations hold: $a_i - a_j = a_i' - a_j'$ and $b_u - b_v = b_u' - b_v'$. In order to prove that this representation is unique, let $M_{ij} = a_i' + b_j'$ and $M_{ij} = a_i'' + b_j''$ be two such representations. Thus $a_i' + b_j' = a_i'' + b_j''$ for $1 \leq i \leq r, 1 \leq j \leq c$. For each $i$, if we sum both sides with respect to $j$, we obtain $ca_i' = ca_i''$, from which we obtain $a_i' = a_i''$ for all $i$. So $a_i' + b_j' = a_i' + b_j'$ for $1 \leq i \leq r, 1 \leq j \leq c$, from which we obtain $b_j' = b_j''$ for all $j$.

Now let us state a formal definition of a two factor design with one observation per cell and no interaction. It is a set of $rc$ observable random variables \{$X_{ij} : 1 \leq i \leq r, 1 \leq j \leq c$\}, $r + c - 1$ numbers \{$a_1, \ldots, a_r, b_1, \ldots, b_{c-1}$\} and $rc$ unobservable random variables \{$Z_{ij} : 1 \leq i \leq r, 1 \leq j \leq c$\} that are independent and identically distributed with common distribution $\mathcal{N}(0, \sigma^2)$, \(\sigma^2 > 0\) unknown, that satisfy

$$X_{ij} = a_i + b_j + Z_{ij},$$

for every $1 \leq i \leq r, 1 \leq j \leq c$.

Note that there are only $c - 1$ $b_j$'s; this is because of the requirement that $b_1 + \cdots + b_c = 0$, which yields

$$b_c = -b_1 - \cdots - b_{c-1}.$$

Thus if we wish to test the null hypothesis that there are no differences among the levels of the first factor, this is the same as testing the null hypothesis $H_0 : a_1 = \cdots = a_r = (\text{some common number}) \ a$. We now put together the components of the general linear model and the linear model when the null hypothesis is true.

First, as we have noted before, the null hypothesis that there are no differences in response among the $r$ levels of the first factor is equivalent to the statement that all numbers in \{$a_1, a_2, \ldots, a_r$\} are the same number, $a$. In other words, the null hypothesis that we want to test is

$$H : a_1 = a_2 = \cdots = a_r = (\text{some}) \ a$$

against the alternative that not every pair of $a_i$'s are equal. Let us first set up the linear model $Y = X\beta + Z$ for these data. First

$$Y = (X_{11} \cdots X_{1c} \ X_{21} \cdots X_{2c} \cdots X_{r1} \cdots X_{rc})',$$

205
which implies that \( n = rc \). Then

\[
\beta = (a_1 \cdots a_r b_1 \cdots b_{c-1})^t
\]

which implies that \( k = r + c - 1 \). We shall spell out the design matrix, \( X \), in some detail. It is an \( n \times (r + c - 1) \) matrix. Its first row is

\[
(1 0 \cdots 0 1 0 \cdots 0),
\]

where the second 1 occurs in the \((r + 1)\)th place. One easily sees that this row matrix postmultiplied by \( \beta \) gives \( a_1 + b_1 \) which is equal to \( E(X_{11}) \). The second row is

\[
(1 0 \cdots 0 1 0 \cdots 0),
\]

where the second 1 occurs in the \((r + 2)\)th place. Again, one easily sees that this row matrix postmultiplied by \( \beta \) gives \( a_1 + b_2 \) which is equal to \( E(X_{12}) \). The \((c - 1)\)th row is

\[
(1 0 \cdots 0 0 \cdots 0 1)
\]

which takes care of \( E(X_{1,c-1}) \). The \( c \)th row is

\[
(1 0 \cdots 0 -1 -1 \cdots -1),
\]

where the last \( c - 1 \) places are occupied by \(-1\)'s, because \( b_c = -b_1 - b_2 - \cdots - b_{c-1} \). In rows \( c + 1 \) through \( 2c \), the first \( r \) entries are

\[
0 1 0 \cdots 0,
\]

and the last \( c - 1 \) columns of rows \( c + 1 \) through \( 2c \) turn out to be the same as the last \( c - 1 \) columns of rows 1 through \( c \). Let us go now to the bottom \( c \) rows of \( X \). The first \( r \) entries of each row are

\[
0 0 \cdots 0 1,
\]

and the last \( c - 1 \) columns of the bottom \( c \) rows of \( X \) is a repetition of the
last \( c - 1 \) columns of the first \( c \) rows of \( X \). The matrix looks like this:

\[
X = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & 0 \\
1 & 0 & 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 0 & 0 & \cdots & 0 & -1 & -1 & -1 & \cdots & -1 \\
0 & 1 & 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 1 & 0 & \cdots & 0 & -1 & -1 & -1 & \cdots & -1 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
\vdots & 0 & 1 & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 1 & -1 & -1 & -1 & \cdots & -1 \\
\end{pmatrix}
\]

The random vector \( Z \) is of course

\[
Z = (Z_{11} \cdots Z_{1c} Z_{21} \cdots Z_{2c} \cdots Z_{r1} \cdots Z_{rc})^t.
\]

For the linear model \( Y = X_0\gamma + Z \), the parameter vector becomes

\[
\gamma = (a \ b_1 \ \cdots \ b_{c-1})^t,
\]

and the design matrix \( X_0 \) is the \( rc \times c \) matrix formed from \( X \) by replacing columns \( 1, \cdots, r \) by one column of all 1’s, \( 1_n \). Thus \( n = rc \), \( k = r + c - 1 \) and \( q = r - 1 \). Thus the test statistic is

\[
F = \frac{\| X\hat{\beta} - X_0\hat{\gamma} \|^2}{k - 1} \frac{n - k}{\| Y - X\hat{\beta} \|^2} = \frac{\| X\hat{\beta} - X_0\hat{\gamma} \|^2}{r - 1} \frac{rc - r - c + 1}{\| Y - X\hat{\beta} \|^2},
\]

or

\[
F = \frac{\| X\hat{\beta} - X_0\hat{\gamma} \|^2}{\| Y - X\hat{\beta} \|^2} (c - 1),
\]

where \( \hat{\beta} = (X^tX)^{-1}X^tY \) and \( \hat{\gamma} = (X_0^tX_0)^{-1}X_0^tY \). This statistic has the \( F_{r-1, r, q} \)-distribution. Again, we observe the value of \( F \), call it \( F_0 \), and we find the probability that a random variable \( W \) with the \( F_{r-1, r, q} \)-distribution.

207
distribution can take a value as extreme as $F_0$, i.e., the probability $P(W \geq F_0)$. If this probability is unbelievably small, we reject the null hypothesis and declare that there is at least one pair of levels of the first factor, call them levels $i$ and $j$, $i \neq j$, such that $a_i \neq a_j$.

It should be noted at this point that there can be more than one observation in a cell. In this case the single observation in $Y$ is replaced by the more than one, and the corresponding rows of the design matrix are repeated as many times as there are observations. An example in section 8.7 shows how this is done.

The problem then arises about what to do if the null hypothesis is rejected and the conclusion is drawn that not all of the $a_i$'s are equal. We now have the same problem as with the single factor design. In this case, the joint distribution of the $r$-dimensional random vector $(\hat{a}_1 \cdots \hat{a}_r)^t$ is multivariate normal with mean vector $(a_1 \cdots a_r)^t$ and a covariance matrix $\sigma^2 C$, where $C$ is the $r \times r$ submatrix of $(X^tX)^{-1}$ determined by the first $r$ rows and the first $r$ columns. If we let $C = (c_{ij})$, then it is easily seen that, if $1 \leq i < j \leq r$, then $\hat{a}_i - \hat{a}_j$ is $N(a_i - a_j, \sigma^2(c_{ii} - 2c_{ij} + c_{jj}))$. From this we obtain that

\[
\frac{\hat{a}_i - \hat{a}_j - (a_i - a_j)}{\sigma \sqrt{c_{ii} - 2c_{ij} + c_{jj}}}
\]

is $N(0, 1)$. Also, from previous results,

\[
\frac{\|Y - X\hat{\beta}\|^2}{\sigma^2}
\]

has the chi-square distribution with $n - r - c + 1$ degrees of freedom. Accordingly, the ratio

\[
\frac{(\hat{a}_i - \hat{a}_j - (a_i - a_j))}{\sqrt{c_{ii} - 2c_{ij} + c_{jj}}} \frac{\|Y - X\hat{\beta}\|^2}{(n - r - c + 1)}
\]

does not depend on $\sigma^2$, and it has the $t$-distribution with $n - r - c + 1$ degrees of freedom. From this we are able to determine the $100(1 - \alpha)$ percent confidence intervals for each of the \(\binom{r}{2}\) differences $a_i - a_j$, where $1 \leq i < j \leq r$. This is done in the same way as it was done in the one way analysis of variance. Namely, find a number $t_0$ so that if $W$ is a random variable with the $t$-distribution with $n - r - c + 1$ degrees of freedom, then
\[ P([W \leq t_0]) = 1 - \alpha/2. \] If
\[
L_{ij} = \hat{a}_i - \hat{a}_j - t_0 \sqrt{c_{ii} - 2c_{ij} + c_{jj}} \sqrt{\frac{\|Y - X\hat{\beta}\|^2}{n - r - c + 1}}
\]
and if
\[
U_{ij} = \hat{a}_i - \hat{a}_j + t_0 \sqrt{c_{ii} - 2c_{ij} + c_{jj}} \sqrt{\frac{\|Y - X\hat{\beta}\|^2}{n - r - c + 1}},
\]
then \( P([L_{ij} \leq a_i - a_j \leq U_{ij}]) = 1 - \alpha. \) As in the previous section, if, for \( \alpha > 0 \) sufficiently small, \( 0 \in (L_{ij}, U_{ij}) \), we shall conclude that \( a_i = a_j. \) If \( 0 < L_{ij} \), then we shall conclude that \( a_i > a_j, \) and if \( U_{ij} < 0, \) we shall conclude that \( a_i < a_j. \)

In conclusion it should be mentioned that one can do a permutation test for this model. In brief, it runs like this. Consider the \( j \)th block, and let \( n_{1j} \) denote the number of observations in the top cell, let \( n_{2j} \) denote the number of observations in the second cell from the top, and so on down to \( n_{rj} \) denoting the number of observations in the bottom cell. Assume the null hypothesis is true, and consider all the \( \sum_{i=1}^{r} n_{ij} \) observations in the \( j \)th block to have the same distribution. Now take a random permutation of these \( \sum_{i=1}^{r} n_{ij} \) observations, placing \( n_{1j} \) of them selected at random without replacement in the top cell, taking \( n_{2j} \) of them selected at random without replacement from those remaining and placing them in the second cell from the top, etc. One does this for \( j = 1, 2, \cdots, c. \) Then one computes \( F \) as defined above, repeats this a large number of times and computes the ratio of the number of times one observes \( F > F_0 \) to the number of repetitions or simulations. This ratio is the empirical \( P \)-value of the permutation test. If it is unbelievably small, then one rejects the null hypothesis that all \( a_i \)'s are equal.

**EXERCISES**

1. Prove: if \( a_1, a_2, \cdots, a_r \) are real numbers such that \( \sum_{i=1}^{r} a_i = 0, \) then they are all equal if and only if they are all equal to zero.
2. Prove that \( \text{rank}(X) = r + c - 1 \) and \( \text{rank}(X_0) = c. \)
3. Verify all the statements made in the penultimate paragraph of this section.

**8.7 The General Two-Factor Design.** The two factor design developed in section 8.6 is what is called a balanced design, since it had exactly
one observation per cell, i.e., for each pair determined by a first factor level and a second factor level. Actually, one could have any number of observations in each cell. If these cell sizes are not equal, this would be what one could call an unbalanced design. The analysis of it is the same, except now each row in the design matrix $X$ in the previous section would be repeated as many times as there are observations in that cell. However, one assumption that we made for that model, that the possible differences between levels of the first factor did not depend on the level of the second factor, might not be true. It might be something that the experimenter would like to verify before testing whether there are any differences among levels of the first factor. This question is addressed in this section.

A proper introduction of this design is the following theorem.

**Theorem 1.** Let $\{\mu_{ij} : 1 \leq i \leq r, 1 \leq j \leq c\}$ be real numbers. Then there exists one and only one representation of $\mu_{ij}$ as

$$\mu_{ij} = a_i + b_j + c_{ij}$$

such that

$$\sum_{v=1}^{c} b_v = \sum_{u=1}^{r} c_{uj} = \sum_{v=1}^{c} c_{iv} = 0$$

for $1 \leq i \leq r, 1 \leq j \leq c$.

*Proof:* Let us denote

$$a_i = \frac{1}{c} \sum_{j=1}^{c} \mu_{ij}, \quad b_j = \frac{1}{r} \sum_{i=1}^{r} \mu_{ij} - \frac{1}{rc} \sum_{i=1}^{r} \sum_{j=1}^{c} \mu_{ij},$$

and $c_{ij} = \mu_{ij} - a_i - b_j$. It is easy to verify that with these definitions,

$$\sum_{v=1}^{c} b_v = \sum_{u=1}^{r} c_{uj} = \sum_{v=1}^{c} c_{iv} = 0.$$

We now show that the representation is unique. Suppose

$$\mu_{ij} = a'_i + b'_j + c'_{ij}$$

such that for all $t$,

$$\sum_{v=1}^{c} b'_v = \sum_{i=1}^{r} c'_{it} = \sum_{j=1}^{c} c'_{tj} = 0.$$
Summing both sides of \( \mu_{ij} = a'_i + b'_j + c'_{ij} \) with respect to \( j \) we obtain

\[
\sum_{j=1}^{c} \mu_{ij} = ca'_i \text{ or } a'_i = a_i,
\]

from which there follows that \( \mu_{ij} = a_i + b'_j + c'_{ij} \). Thus \( a_i + b_j + c_{ij} = a_i + b'_j + c'_{ij} \) or \( b_j + c_{ij} = b'_j + c'_{ij} \). Summing both sides with respect to \( i \), we obtain \( b_j = b'_j \) and finally, \( c_{ij} = c'_{ij} \) for \( 1 \leq i \leq r, 1 \leq j \leq c \).

The above theorem provides us with a parameterization of the general two-factor design. Again the linear model is \( \mathbf{Y} = X \mathbf{\beta} + \mathbf{Z} \). The parameter vector, \( \mathbf{\beta} \), must now be represented as

\[
\mathbf{\beta} = (a_1, \cdots, a_r, b_1, \cdots, b_{c-1}, c_{11}, \cdots, c_{1,c-1}, \cdots, c_{r-1,1}, \cdots, c_{r-1,c-1}),
\]

and the parameters \( c_{11}, \cdots, c_{1,c-1}, \cdots, c_{r-1,1}, \cdots, c_{r-1,c-1} \) are called the interaction parameters. We see that

\[
k = r + (c - 1) + (r - 1)(c - 1) = rc.
\]

Thus, in order to be able to satisfy the requirement that \( k < n \), we need at least two observations in at least one cell. From what we have done before, the layout of the design matrix should be rather obvious. Let us suppose that the \( \mathbf{Y} \) vector is laid out from observations for the \( c \) different levels of the first level of the first factor, followed by the same for the second level of the first factor, etc. The explanation of how this design matrix is laid out is too long, but it can easily be illustrated so as to provide everyone with the ability to lay it out. Suppose we have two levels of the first factor and three levels of the second factor. The data are listed as below. There are two cells with two observations.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>21</td>
<td>24</td>
<td>28</td>
<td>30</td>
</tr>
<tr>
<td>B</td>
<td>22</td>
<td>26</td>
<td>27</td>
<td>31</td>
</tr>
</tbody>
</table>

Suppose we wanted to test the null hypothesis that the interaction terms \( c_{11} \) and \( c_{12} \) are both zeros. Then \( \mathbf{Y}, \mathbf{X}, \mathbf{\beta}, X_0 \) and \( \gamma_0 \) are as follows: \( \mathbf{Y}, \mathbf{X} \) and
\( \beta \) are
\[
Y = \begin{pmatrix}
21 \\
24 \\
28 \\
30 \\
22 \\
26 \\
27 \\
31 \\
\end{pmatrix},
\quad X = \begin{pmatrix}
1 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 \\
\end{pmatrix},
\quad \beta = \begin{pmatrix}
a_1 \\
a_2 \\
b_1 \\
b_2 \\
c_{11} \\
c_{12} \\
\end{pmatrix},
\]
and \( X_0 \) and \( \gamma_0 \) are
\[
X_0 = \begin{pmatrix}
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 \\
0 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 \\
0 & 1 & -1 & -1 \\
\end{pmatrix}
\quad \text{and} \quad \gamma = \begin{pmatrix}
a_1 \\
a_2 \\
b_1 \\
b_2 \\
\end{pmatrix}.
\]
In this case, \( n = 8, k = 6 \) and \( q = 2 \). Note that the design matrix under the null hypothesis, \( X_0 \), is obtained from the design matrix \( X \) by deleting the last two columns.

This representation allows us to test whether the model considered in section 8.6 is reasonable. Namely, if in the special model given above, we do not reject the null hypothesis that \( c_{11} = c_{12} = 0 \), then we may be permitted to test whether there is a difference between the two levels of the first factor. If at least one of these interaction terms is not zero, then it is rather meaningless to test for a difference in the row effects.

**EXERCISES**

1. Prove that the design matrix for the two factor design has full rank, i.e., \( \operatorname{rank}(X) = rc \).
2. In the example given in this section, suppose there that the number of observations in the cell determined by the \( i \)th level of the first factor and the \( j \)th level of the second factor is \( n_{ij} \). Derive a \( t \)-test for testing that
8.8 A test for Independence. The test for independence that we shall develop here involves a sample of size $n$ from a $\mathcal{N}_2(\mu, C)$ population. Another way of saying this is to consider a sequence of $n$ observable bivariate random vectors,

$$
\begin{pmatrix}
X_1 \\
Y_1
\end{pmatrix}, \ldots, 
\begin{pmatrix}
X_n \\
Y_n
\end{pmatrix},
$$

which are independent and identically distributed with common distribution $\mathcal{N}_2(\mu, C)$. These are $n$ pairs of measurements on $n$ individuals selected from some population whose distribution is that just given. The problem here is to test the null hypothesis that $X_i$ and $Y_i$ are independent. This occurs when one is measuring two characteristics of each of $n$ individuals, and the question arises whether these two responses are independent of each other.

We begin by recalling corollary 2 of theorem 8 in Chapter 7. This theorem states that if a two dimensional random vector $(X, Y)^t$ is $\mathcal{N}_2(\mu, C)$, then there exist unique constants $a$ and $b$ and a random variable $Z$ such that $Y = a + bX + Z$, where $X$ and $Z$ are independent and $E(Z) = 0$. From this it easily follows that $a = E(Y) - bE(X)$ and

$$
b = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{s.d.(Y)}{s.d.(X)}\rho_{X,Y},
$$

where $\rho_{X,Y}$ denotes the correlation coefficient. Corollary 1 to theorem 7 in section 7.5 implies that $X$ and $Y$ are independent if and only if $\text{Cov}(X, Y) = 0$, i.e., if and only if $b = 0$. Thus we wish to obtain a test of the null hypothesis that $b = 0$.

The basis of the test to be proposed is the next theorem whose proof is beyond the scope of this course. Let us suppose that we have the sample of $n$ random vectors introduced above that are independent and have common distribution $\mathcal{N}_2(\mu, C)$. First we define the random vector $Y$, the $n \times 2$ random matrix $X(X)$ and the $n \times 1$ matrix by

$$
Y = \begin{pmatrix}
Y_1 \\
\vdots \\
Y_n
\end{pmatrix},
X(X) = \begin{pmatrix}
1 & X_1 \\
\vdots & \vdots \\
1 & X_n
\end{pmatrix}
\quad \text{and} \quad X_0 = 1_n.
$$

Then we prove the following lemma.
Lemma 1. With probability one, the rank of the matrix $X(X)$ is 2.

Proof: It is sufficient to prove that the determinant of the matrix determined by the first two rows is unequal to zero with probability one. But this determinant is $X_2 - X_1$. This random variable has a normal distribution, and thus its distribution function is absolutely continuous. This in turn implies that the probability that $X_2 - X_1$ equals any number, in particular, 0, is zero.

With lemma 1 it is possible to write $(X(X)^t X(X))^{-1}$, which exists with probability one. Let us now define the observable random vectors

$$\widehat{\beta}(X, Y) = (X(X)^t X(X))^{-1} X(X)^t Y$$ and $$\widehat{\gamma}(Y) = (X_0^t X_0)^{-1} X_0 Y,$$

and finally define

$$F(X) = \frac{\| X(X)\widehat{\beta}(X, Y) - X_0 \widehat{\gamma}(Y) \|^2 n - 2}{\| Y - X(X)\widehat{\beta}(X, Y) \|^2}.$$

Theorem 1. Under the above notation and hypotheses, if $X_1$ and $Y_1$ are independent, then the observable random variable $F(X)$ has the $F$-distribution with $(1, n - 2)$ degrees of freedom.

As was mentioned above, the proof of this theorem is beyond the scope of this course. The statistic $F(X)$ can be used as a test for independence. Simply put, observe its value. If it is small there is no reason to doubt the null hypothesis. But if the observed value $F_0$ of it is unbelievably large, i.e., if the probability that a random variable with the $F_{1, n-2}$-distribution is greater that $F_0$ is unbelievably small, then we would reject the null hypothesis and state that they are not independent.

EXERCISES

1. Provide another proof of lemma 1 based on linear independence of the columns.