1 Least squares fits

This section has no probability in it. There are no random variables. We are given n points (x_i, y_i) and want to find the equation of the line that best fits them. We take the equation of the line to be

$$y = \alpha + \beta x \tag{1}$$

We have to decide what "best fit" means. For each point the value of the line at x_i will be $\alpha + \beta x_i$. So the vertical distance between the line and the actual data point (x_i, y_i) is $|y_i - \alpha - \beta x_i|$. We take "best fit" to mean the choice of α and β that minimizes the sum of the squares of these errors:

$$Q = \sum_{i=1}^{n} [y_i - \alpha - \beta x_i]^2 \tag{2}$$

To find the minimum we find the critical points: take the partial derivatives of this with respect to α and β and set them to zero.

$$\begin{aligned} 0 &= \frac{\partial Q}{\partial \alpha} = 2 \sum_{i=1}^{n} [y_i - \alpha - \beta x_i](-1) \\ 0 &= \frac{\partial Q}{\partial \beta} = 2 \sum_{i=1}^{n} [y_i - \alpha - \beta x_i](-x_i) \end{aligned} \tag{3}$$

Define

$$\begin{aligned} \overline{X} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ \overline{Y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\ \overline{X^2} &= \frac{1}{n} \sum_{i=1}^{n} x_i^2 \\ \overline{XY} &= \frac{1}{n} \sum_{i=1}^{n} x_i y_i \end{aligned} \tag{4}$$

Then solving the above two equation for α and β gives (after some algebra)

$$\begin{aligned} \beta &= \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - (\overline{X})^2} \\ \alpha &= \overline{Y} - \beta\overline{X} \end{aligned} \tag{5}$$

2 Simple linear regression

Suppose we have some experiment for which we can set the value of an input x and them measure some output y. For example, we put a metal rod in an oven. x is the temperature we set the oven to and y is the length of the rod. (Metals usually expand when heated.) We believe that y is a linear function of x. However, because of experimental error (or other sources of noise) the measured value of y is not a linear function of x.

We model this situation by something called a "simple linear model". Let x_i, $i = 1, 2, \cdots, n$ be the values of x used in our experiment. We let

$$Y_i = \alpha + \beta x_i + \epsilon_i \tag{6}$$

Where the ϵ_i are iid normal random variables with mean zero and common variance σ^2. The three parameters α, β, σ are unknown. We are given n data points (x_i, Y_i), and our job is to estimate the three parameters or test hypotheses involving them. Keep in mind that in this model the x_i are non random, the Y_i are random.

The above equation implies the Y_i will be independent normal random variables with mean $\alpha + \beta x_i$ and variance σ^2. Thus the joint density of the Y_i is

$$f(y_1, y_2, \cdots, y_n|\alpha, \beta, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2] \tag{7}$$

We will do maximimum likelihood estimation. So given $y_1, \cdots, y_n$ (and $x_1, \cdots, x_n$), we want to maximize the likelihood function as a function of α, β, σ. The natural thing to do is to take the partial derivatives with respect to each of the three parameters, set them all to zero and do a bunch of algebra. We can save some algebra by the following approach. First we think of σ as fixed and maximize f over α and β. This is equivalent to minimizing

$$\sum_{i=1}^{n} [y_i - \alpha - \beta x_i]^2 \tag{8}$$

But this is exactly the problem we solved in the last section. So the optimal α and β are given by equation 5. Note that they do not depend on σ. Denoting the above quantity by Q we must now maximize

$$\frac{1}{(2\pi\sigma^2)^{n/2}}\ \exp[-\frac{1}{2\sigma^2}Q] \tag{9}$$

It is a bit easier to minimize the log of this likelihood function.

$$\frac{\partial}{\partial\sigma}\ln f = \frac{\partial}{\partial\sigma}\left(-n\ln\sigma - \frac{1}{2\sigma^2}Q\right) = \frac{-n}{\sigma} + \frac{1}{\sigma^3}Q \tag{10}$$

Setting this to zero leads to

$$\sigma^2 = \frac{1}{n}Q \tag{11}$$

We have found the maximum likelihood estimators. They are

$$\begin{aligned}
\hat{\beta} &= \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - (\overline{X})^2} \\
\hat{\alpha} &= \overline{Y} - \hat{\beta}\overline{X} \\
\hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2
\end{aligned} \tag{12}$$

$\overline{X}$,$\overline{Y}$, ... are defined as in the previous section with y_i replaced by Y_i.

The estimators are random variables since then involve the random variables Y_i. An important question is how good are they? In particular, are they unbiased and what is their variance.

Each of the estimators α and β can be written as linear combinations of the Y_i.

$$\hat{\beta} = \sum_{j=1}^{n} c_j Y_j \tag{13}$$

It is convenient to define

$$s_X^2 = \sum_{i=1}^{n}(x_i - \overline{X})^2 = n[\overline{X^2} - (\overline{X})^2] \tag{14}$$

(The last equality is easily checked with a little algebra.) Then some algebra shows

$$c_j = \frac{x_j - \overline{X}}{s_X^2} \tag{15}$$

Thus

$$E[\hat{\beta}] = \sum_{j=1}^{n} c_j E[Y_j] = \sum_{j=1}^{n} c_j(\alpha + \beta x_i) = \beta \tag{16}$$

The last equality follows from

$$\sum_{j=1}^{n} c_j = 0, \quad \sum_{j=1}^{n} c_j x_j = 1 \tag{17}$$

Since the Y_j are independent, we also get the variance:

$$var(\hat{\beta}) = \sum_{j=1}^{n} c_j^2 \, var(Y_j) = \sum_{j=1}^{n} c_j^2 \, \sigma^2 = \frac{\sigma^2}{s_X^2} \tag{18}$$

Similarly, we write α as a linear combination of the Y_i.

$$\hat{\alpha} = \sum_{j=1}^{n} d_j Y_j \tag{19}$$

where some algebra shows

$$d_j = \frac{\overline{X^2} - x_j \overline{X}}{s_X^2} \tag{20}$$

Thus

$$E[\hat{\alpha}] = \sum_{j=1}^{n} d_j E[Y_j] = \sum_{j=1}^{n} d_j(\alpha + \beta x_i) = \alpha \tag{21}$$

The last equality follows from

$$\begin{aligned} \sum_{j=1}^{n} d_j = \frac{n\overline{X^2} - n(\overline{X})}{s_X^2} &= 1 \\ \sum_{j=1}^{n} c_j x_j &= 0 \end{aligned} \tag{22}$$

Since the Y_j are independent, we also get the variance:

$$var(\hat{\alpha}) = \sum_{j=1}^{n} d_j^2 \, var(Y_j) = \sum_{j=1}^{n} d_j^2 \, \sigma^2 = \frac{\sigma^2 \, \overline{X^2}}{s_X^2} \tag{23}$$

3 Joint distribution of the estimators

The goal of this section is to find the joint distribution of the estimators $\hat{\alpha}, \hat{\beta}$ and $\hat{\sigma^2}$. The key ingredient is the following theorem. Recall that an n by n matrix A is orthogonal if $A^t A = I$ where A^t is the transpose of A. Orthogonal matrices preserve lengths of vectors: $||Ax|| = ||x||$ for $x \in R^n$ if A is orthogonal. A matrix is orthogonal if and only if its rows form an orthonormal basis of R^n. A matrix is orthogonal if and only if its columns form an orthonormal basis of R^n.

Theorem 1. *Let $Y_1, Y_2, \cdots, Y_n$ be independent, normal random variables with the same variance σ^2. (Their means are arbitrary.) Let A be an n by n orthogonal matrix. Let $Z = AY$ and let Z_i be the components of Z, so $Z = (Z_1, Z_2, \cdots, Z_n)$. Then $Z_1, Z_2, \cdots, Z_n$ are independent normal random variables, each with variance σ^2.*

We apply this theorem to our random variables Y_i and the following orthogonal matrix A. The first row of A is

$$a_{1j} = \frac{1}{\sqrt{n}}, \quad j = 1, 2, \cdots, n$$

The second row is

$$a_{2j} = \frac{x_j - \overline{X}}{s_X}, \quad j = 1, 2, \cdots, n$$

where s_X is defined as before:

$$s_X^2 = \sum_{i=1}^{n} (x_i - \overline{X})^2$$

It is easy to check that these two rows form an orthonormal set. (They are orthogonal and each has norm equal to one.) By the Gram-Schmidt process we can find $n-2$ more vectors which together with the first two rows form an orthonormal basis. We use these $n-2$ vectors as the remaining rows. This gives us an orthogonal matrix. As in the theorem we define $Z = AY$.

We will express the estimators for α, β, σ^2 in terms of the Z_i. We have

$$Z_1 = \sum_{j=1}^{n} \frac{1}{\sqrt{n}} Y_j = \sqrt{n}\overline{Y}$$

$$Z_2 = \frac{1}{s_X} \sum_{j=1}^{n} (x_j - \overline{X}) Y_j = \frac{n}{s_X} (\overline{XY} - \overline{X}\,\overline{Y})$$

Thus we have

$$\hat{\beta} \;=\; \frac{\overline{XY} - \overline{X}\,\overline{Y}}{s_X^2/n} = \frac{1}{s_X} Z_2$$

And we have

$$\hat{\alpha} \;=\; \overline{Y} - \beta \overline{X} = \frac{1}{\sqrt{n}} Z_1 - \frac{\overline{X}}{s_X} Z_2 \tag{24}$$

Since A is orthogonal,

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} Z_i^2$$

A fair amount of algebra shows that if we let

$$S^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \tag{25}$$

then

$$S^2 = \sum_{i=3}^{n} Z_i^2 \tag{26}$$

Note that the sum starts at $i = 3$. The Z_i are independent normal random variables, each with variance σ^2. It is not hard to show that the mean of Z_i is zero if $i \geq 3$. So S^2/σ^2 has a χ^2 distribution with $n-2$ degrees of freedom. In particular, the mean of S^2 is $(n-2)\sigma^2$.

The maximum likelihood estimator for σ^2 is S^2/n. Thus we have now found its mean:

$$E[\hat{\sigma^2}] = \frac{n-2}{n} \sigma^2 \tag{27}$$

So the maximum likelihood estimator is biased. The corresponding unbiased estimator is $S^2/(n-2)$. The estimators $\hat{\alpha}, \hat{\beta}$ are linear combinations of Z_1 and Z_2, while $\hat{\sigma^2}$ only depends on Z_i for $i \geq 3$. Thus $\hat{\alpha}$ and $\hat{\beta}$ are independent of $\hat{\sigma^2}$. We summarize our results in the following theorem

Theorem 2. *$\hat{\alpha}$ and $\hat{\beta}$ have a bivariate normal distribution with*

$$\begin{aligned} E[\hat{\alpha}] &= \alpha, \quad E[\hat{\beta}] = \beta \\ var(\hat{\alpha}) &= \frac{\overline{X^2}\sigma^2}{s_X^2}, \quad var(\hat{\beta}) = \frac{\sigma^2}{s_X^2} \\ Cov(\hat{\alpha}, \hat{\beta}) &= E[\hat{\alpha}\hat{\beta}] - E[\hat{\alpha}]E[\hat{\beta}] = -\frac{\overline{X}\sigma^2}{s_X^2} \end{aligned} \tag{28}$$

S^2 (and hence $\hat{\sigma^2}$) is independent of $\hat{\alpha}, \hat{\beta}$. S^2/σ^2 has a χ^2 distribution with $n-2$ degrees of freedom.

We can use the theorem to do hypothesis testing involving α and β and to find confidence intervals for them.

We start with hypothesis testing involving β. Consider

$$\begin{aligned} H_0 : &\quad \beta = \beta^* \\ H_1 : &\quad \beta \neq \beta^* \end{aligned} \tag{29}$$

where β^* is a constant. We have taken the alternative to be two sided, but the case of a one-sided alternative is similar. If the null hypothesis is true, then by the theorem,

$$\frac{\hat{\beta} - \beta^*}{\sigma/s_X} = \frac{\hat{\beta} - \beta}{\sigma/s_X} = s_X \frac{\hat{\beta} - \beta}{\sigma} \tag{30}$$

has a standard normal distribution. Since σ is unknown, we must estimate it. So we define the statistic to be

$$T = s_X \frac{\hat{\beta} - \beta^*}{\sqrt{S^2/(n-2)}} \tag{31}$$

Note that we have used the unbiased estimator of σ^2. We can write the above as

$$\frac{U}{\sqrt{V/(n-2)}} \tag{32}$$

where

$$U = \frac{\hat{\beta} - \beta^*}{\sigma / s_X} \tag{33}$$

and

$$V = \frac{S^2}{\sigma^2} \tag{34}$$

This shows that T has a student-t distribution with $n-2$ degrees of freedom.

Now given a significance level ϵ, we choose c so that

$$P(|T| \geq c) = \epsilon \tag{35}$$

Then the test is to reject the null hypothesis if $|T| \geq c$.

For confidence intervals, suppose we want a 95% confidence interval. Then we choose a (using tables or software) so that

$$P(|T| \leq a) = 0.95 \tag{36}$$

Then the confidence interval is

$$[\hat{\beta} - \frac{a}{s_X}\sqrt{S^2/(n-2)}, \hat{\beta} + \frac{a}{s_X}\sqrt{S^2/(n-2)}] \tag{37}$$

For hypothesis testing on α, consider

$$\begin{aligned} H_0 : &\quad \alpha = \alpha^* \\ H_1 : &\quad \alpha \neq \alpha^* \end{aligned} \tag{38}$$

Taking into account the variance of $\hat{\alpha}$, we see that the approriate statistic is now

$$T = s_X \frac{\hat{\alpha} - \alpha^*}{\sqrt{\overline{X^2} S^2/(n-2)}} \tag{39}$$

It has a student's t distribution with $n-2$ degrees of freedom. The confidence interval now has the form

$$[\hat{\alpha} - \frac{a}{s_X}\sqrt{\overline{X^2} S^2/(n-2)}, \hat{\alpha} + \frac{a}{s_X}\sqrt{\overline{X^2} S^2/(n-2)}] \tag{40}$$

4 General linear regression

The model is now

$$Y_i = \sum_{j=1}^{k} X_{ij}\beta_j + \epsilon_i \tag{41}$$

where X is a n by k matrix. The β_j are k unknown parameters and the ϵ_i are iid normal random variables with mean zero and variance σ^2. The matrix X is sometimes called the *design matrix*. We assume it has trivial null space, i.e., the only k-dimensional vector z such that $Xz = 0$ is $z = 0$. (This is equivalent to the columns of X being independent.)

This models includes several interesting special cases. We will consider one of them (ANOVA) later. For now we point out the following special case. In the previous section we assumed that y was a linear function of x plus some "noise." Now suppose that y is an mth degree polynomial function of x plus some noise. The coeffecients of the polynomial are unknown. So the model should be

$$Y_i = \sum_{j=0}^{m} \beta_j x_i^j + \epsilon_i \tag{42}$$

where $i = 1, 2, \cdots, n$. If we let $k = m + 1$, let

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots x_1^m \\ 1 & x_2 & x_2^2 & \cdots x_2^m \\ \cdots & & & \\ 1 & x_n & x_n^2 & \cdots x_n^m \end{pmatrix}, \tag{43}$$

let β be the vector $(\beta_0, \beta_1, \cdots, \beta_m)$ and let $\epsilon = (\epsilon_1, \cdots, \epsilon_n)$, then $Y = X\beta + \epsilon$. So our polynomial regression is a special case of the general linear model.

To find estimators for the unknown parameters β_j and σ^2 we use the maximum likelihood estimators. The joint distribution of the Y_i is

$$f(y_1, \cdots, y_n|\beta, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{1}{2\sigma^2}Q] \tag{44}$$

where

$$Q = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} X_{ij}\beta_j)^2 \tag{45}$$

Again, it is convenient first fix σ and maximize f over all the β_j. This is equivalent to minimizing Q. This is an interesting linear algebra problem, whose solution may be found in the appendix on linear algebra in the book. Let y denote the n dimensional vector $(y_1, \cdots, y_n)$, β the k dimensional vector $(\beta_1, \cdots, \beta_n)$. Then we can write Q as $||y - X\beta||^2$. The set of all vectors of the form $X\beta$ is a subspace of R^n. If we assume that the columns of X are linearly independent (X has rank k), then the set of vectors of the form $X\beta$ is a k dimensional subspace of R^n. Minimizing Q is equivalent to finding the vector in this subspace that is closest to y. A theorem in linear algebra says this is given by

$$X^t X \hat{\beta} = X^t y \tag{46}$$

The assumption that the rank of X is k implies that the k by k matrix $X^t X$ is invertible, so

$$\hat{\beta} = (X^t X)^{-1} X^t y \tag{47}$$

Finally we minimize the likehood function (with this choice of β) over σ. This gives the following estimator for σ^2:

$$\hat{\sigma^2} = \frac{||Y - X\hat{\beta}||^2}{n} \tag{48}$$

Since ϵ_i has mean 0, the expectation of Y_i is

$$E[Y_i] = \sum_{j=1}^{k} X_{ij}\beta_j \tag{49}$$

Or in matrix notation, $E[Y] = X\beta$. Thus

$$E[\hat{\beta}] = E[(X^t X)^{-1} X^t Y] = (X^t X)^{-1} X^t E[Y] = (X^t X)^{-1} X^t X\beta = \beta \tag{50}$$

Thus the estimators of the β_j are unbiased.

It turns out that the expected value of $||Y - X\hat{\beta}||^2$ is $(n-k)\sigma^2$. So the maximum likelihood estimator of σ^2 is biased. A possible unbiased estimator is

$$\frac{||Y - X\hat{\beta}||^2}{n-k} \tag{51}$$

which is the estimator given in the book (eq. (14.13)).

The estimators $\hat{\beta}$ are linear combinations of the independent normal RV's Y_i. This implies that their joint distribution is a multivariate normal. You probably haven't seen this. One property of a multivariate normal distribution is that any linear combination of the RV's will have a normal distribution. In particular, each $\hat{\beta}_i$ has a normal distribution. We have already seen that its mean is β_i, so if we knew its variance we would known its distribution completely. The following theorem addresses this. The covariance matrix of $\hat{\beta}$ is the k by k matrix defined by

$$C_{ij} = cov(\hat{\beta}_i, \hat{\beta}_j) \tag{52}$$

Theorem 3. *The covariance matrix of $\hat{\beta}$ is*

$$C = \sigma^2 \, (X^t X)^{-1} \tag{53}$$

In particular, the variance of $\hat{\beta}_i$ is σ^2 times the ith diagonal entry of the matrix $(X^t X)^{-1}$.

Define

$$S^2 = ||Y - X\hat{\beta}||^2 = \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{k} X_{ij} \hat{\beta}_j)^2 \tag{54}$$

Recall that the maximum likelihood estimator of σ^2 was S^2/n. It can be shown that S^2 is independent of all the $\hat{\beta}_i$ and furthermore that S^2/σ^2 has a χ^2 distribution with $n - k$ degrees of freedom.

We can now test hypotheses that involve a single β_i and compute confidence intervals for a single β_i. Fix an index i and consider the null hypothesis $H_0 : \beta_i = \beta^*$ where β^* is a constant. If the null hypothesis is true, then

$$\frac{\hat{\beta}_i - \beta_i^*}{\sqrt{var(\hat{\beta}_i)}} \tag{55}$$

has a standard normal distribution. Let d_i^2 be the ith diagonal entry in $(X^t X)^{-1}$. So the variance of $\hat{\beta}_i$ is $d_i^2 \sigma^2$, and the above becomes

$$\frac{\hat{\beta}_i - \beta_i^*}{d_i \sigma} \tag{56}$$

We do not know σ, so we replace it by the unbiased estimator, $S/\sqrt{n-k}$. So we use the statistic

$$T = \frac{(\hat{\beta}_i - \beta_i^*)\sqrt{n-k}}{d_i S} \tag{57}$$

As before we can rewrite this as

$$T = \frac{(\hat{\beta}_i - \beta_i^*)}{d_i \sigma} \frac{\sigma\sqrt{n-k}}{S} \tag{58}$$

to see it has a student-t distribution with $n-k$ degrees of freedom. Hypothesis testing and confidence intervals then go in the usual way.

5 The F Distribution

Let Y and W be independent random variables such that Y has a χ^2 distribution with m degrees of freedom and W has a χ^2 distribution with n degrees of freedom. (m and n are positive integers.) Define

$$X = \frac{Y/m}{W/n} \tag{59}$$

Then the distribution of X is called the *F distribution with m and n degrees of freedom.* It is possible to explicitly compute the p.d.f. of this distribution, but we will not do so.

Our main interest in the F distribution in the application in the next section. Here we will give a simple application.

Suppose we have two normal populations, one with mean μ_1 and variance σ_1^2 and the other with mean μ_2 and variance σ_2^2. All four of these parameters are unknown. We have a random sample $X_1, \cdots, X_m$ from population 1 and a random sample $Y_1, \cdots, Y_n$ from population 2. The two samples are independent. We want to test the hypotheses:

$$\begin{aligned} H_0 : &\quad \sigma_1^2 \le \sigma_2^2 \\ H_1 : &\quad \sigma_1^2 > \sigma_2^2 \end{aligned} \tag{60}$$

Define

$$S_X^2 = \sum_{i=1}^{m} (X_i - \overline{X_m}), \quad S_Y^2 = \sum_{i=1}^{n} (Y_i - \overline{Y_n}), \tag{61}$$

where $\overline{X_m}$ is the sample mean for $X_1, \cdots, X_m$ and $\overline{Y_n}$ is the sample mean for $Y_1, \cdots, Y_n$. Let

$$T = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)} \tag{62}$$

Then the test is that we should reject H_0 if T is large. It can be shown that if $\sigma_1^2 = \sigma_2^2$, then T has an F distribution with $m-1$ and $n-1$ degrees of freedom. If we want a significance level of α, then we choose c so that for this distribution, $P(T > c) = \alpha$. Then we reject H_0 if $T > c$.

6 Analysis of Variance (ANOVA)

We consider a problem known as the "one way layout." There are p different populations, each has a possibly different mean μ_i, but they all have the same variance σ^2. For $i = 1, 2, \cdots, p$ we have a random sample with n_i observations from the ith population. We denote it by $Y_{i1}, Y_{i2}, \cdots, Y_{in_i}$. We let $n = \sum_{i=1}^p n_i$ be the total number of observations. We want to test the hypothesis that the means of the populations are all the same.

Example

The model is

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \cdots, p, \quad j = 1, \cdots, n_i \tag{63}$$

where the ϵ_{ij} are iid normal random variables with mean zero and variance σ^2. This is a special case of the general linear model. Here the unknown parameters are the μ_i and σ^2.

We define the vectors

$$Y = (Y_{11}, \cdots, Y_{1n_1}, Y_{21}, \cdots, Y_{2n_2}, \cdots, Y_{p1}, \cdots, Y_{pn_p}) \tag{64}$$

$$\epsilon = (\epsilon_{11}, \cdots, \epsilon_{1n_1}, \epsilon_{21}, \cdots, \epsilon_{2n_2}, \cdots, \epsilon_{p1}, \cdots, \epsilon_{pn_p}) \tag{65}$$

$$\mu = (\mu_1, \cdots, \mu_p) \tag{66}$$

The design matrix X, which is n by p, contains only 1's and 0's, with exactly one 1 in each row. Each column in the matrix corresponds to one of the populations. The first n_1 rows have a 1 in the first column. The next n_2 rows have a 1 in the second column. The next n_3 rows have a 1 in the third column. And so on. So the model in matrix notation is $Y = X\mu + \epsilon$.

We can use the results for the general linear model to find the maximum likelihood estimators for the μ_i and their variances. We need to compute $(X^tX)^{-1}$. A little thought shows that X^tX is just a diagonal matrix with $n_1, n_2, \cdots, n_p$ along the diagonal. So its inverse is the diagonal matrix with $1/n_1, 1/n_2, \cdots, 1/n_p$ along the diagonal. And we see that the ith entry of X^tY is just the sum of the observations in the ith sample. We define $\overline{Y_{i+}}$ to be the mean of the sample from population i. So

$$\overline{Y_{i+}} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \tag{67}$$

Thus eq. (47) says that the maximum likelihood estimator for μ_i is just $\overline{Y_{i+}}$. And by the theorem in the previous section, the variance of $\hat{\beta}$ is σ^2 times the ith diagonal entry of $(X^t)^{-1}$, i.e., it is σ^2/n_i. These results are what you would expect if you had never seen the general linear model. If all we want to study is the parameter μ_i for population i, then we can forget about the samples from the other populations and just use the sample from population i. Then we are back to a problem we considered when we first did estimation and hypothesis testing.

The maximum likelihood estimator of σ^2 is

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{i+}})^2 \tag{68}$$

We now turn to the problem we are really interested in, testing the hypotheses:

$$\begin{aligned} H_0: &\qquad \mu_1 = \mu_2 = \cdots \mu_p \\ H_1: &\qquad H_0 \textit{ not true} \end{aligned} \tag{69}$$

We define $\overline{Y_{++}}$ to be the average of all the random samples:

$$\overline{Y_{++}} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{p} n_i \overline{Y_{i+}} \tag{70}$$

We define

$$S^2_{tot} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{++}})^2 \tag{71}$$

The subscript *tot* stands for total. If the null hypothesis is true, then S^2_{tot}/n would be the MLE of σ^2.

Define

$$S^2_{resid} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y_{i+}})^2 \tag{72}$$

$$S^2_{betw} = \sum_{i=1}^{p} n_i (\overline{Y_{i+}} - \overline{Y_{++}})^2 \tag{73}$$

Here *resid* and *betw* stand for residual and between. Some algebra shows that

$$S_{tot}^2 = S_{resid}^2 + S_{betw}^2 \tag{74}$$

Since the p random samples are independent, S_{resid}^2 is the sum of p independent random variables, each of which has a χ^2 distribution with $n_i - 1$ degrees of freedom. Hence S_{resid}^2 has a χ^2 distribution with $\sum_{i=1}^p (n_i - 1) = n - p$ degrees of freedom. Furthermore, since S_{betw}^2 only depends on the random samples through their sample means, S_{betw}^2 is independent of S_{resid}^2. It can also be shown that S_{betw}^2 has χ^2 distribution with $p - 1$ degrees of freedom. Thus we have partitioned the total variation of all the samples the sum of two independent terms - one is the sum of the variations of each sample around its mean and the other reflects how much these sample means vary around the mean of all the samples together.

Now we define the statistic we will use to test our hypotheses:

$$U^2 = \frac{S_{betw}^2/(p-1)}{S_{resid}^2/(n-p)} \tag{75}$$

If the null hypothesis H_0 is true, then U^2 has a F distribution with $p-1$ and $n-p$ degrees of freedom. A large value of U^2 indicates the null hypothesis is not true. Given a significance level α, we pick c so that for the F distribution with $p-1$ and $n-p$ degrees of freedom, $P(U^2 > c) = \alpha$. Then the test is to reject the null hypothesis if $U^2 > c$. The value of c can be found in tables, or better yet from software.