Chapter 5

Distribution Theory

In this chapter, we summarize the distributions related to the normal distribution that occur in linear models. Before turning to this general problem that assumes normal errors, we discuss asymptotic distributions, assuming only the two-moment assumptions on the errors discussed previously in these notes.

5.1 Consistency of least squares estimates

We begin with a few consistency results that stand on their own and do not depend on normality.

Definition 5.1 (Weak consistency) *Suppose that θ_n is an estimate of a parameter θ based on a sample of size n. We say θ_n is a weakly consistent estimate of θ if* $\lim \Pr(|\theta_n - \theta| > \delta) = 0$ *for any $\delta > 0$, or equivalently* $\lim \mathit{Var}(\theta_n) = 0$.

Consider now the standard model:

$$y = \mu, \mu \in \mathcal{E} = \mathcal{R}(X) \tag{5.1}$$

This model is commonly written in the *additive error form*, with

$$y = \mu + \varepsilon, \mu \in \mathcal{E} = \mathcal{R}(X), \mathrm{E}(\varepsilon) = 0, \mathrm{Var}(\varepsilon) = \sigma^2 I \tag{5.2}$$

Models (5.1) and (5.2) are equivalent formulations. While not all statistical models can be formulated using additive errors (for example, this doesn't make any sense for binomial errors), this formulation makes some sense for linear models. This model includes only two moment assumptions concerning ε, but does not give the exact distribution. We know that $\hat{\mu} = Py$, and $\mathrm{E}(\hat{\mu}) = \mu$ and $\mathrm{Var}(\hat{\mu}) = \sigma^2 P$. Suppose we allow n to increase. As n increases in the above sequences, the vector space keeps expanding, and $\mathcal{E}$, as a consequence, changes for each n.

Theorem 5.1 (Huber (1981, p. 157) *Under (5.1), if the errors are independent, then:*

1. *$\hat{\mu}_i$ is a consistent estimator of μ_i if and only if $h_{ii} \to 0$, where h_{ii} is the i-th diagonal element of the projection matrix P.*

2. *$\hat{\mu}$ is a consistent estimate of μ if and only if $\max h_{ii} \to 0$.*

Proof. First, assume that $h_{ii} \to 0$. To demonstrate weak consistency, we need to show that:

$$\Pr(|\hat{\mu}_i - \mu_i| > \delta) \to 0$$

as $n \to \infty$. But $\hat{\mu}_i$ is unbiased for μ_i, and thus by the Tchebychev inequality:

$$\Pr(|\hat{\mu}_i - \mu_i| > \delta) \leq \frac{\mathrm{var}(\hat{\mu}_i)}{\delta^2} = \frac{h_{ii}\sigma^2}{\delta^2}$$

and the conclusion follows. Independence is not used here so this part is true for uncorrelated errors also.

To show necessity, we use contradiction, and bound $\Pr(|\hat{\mu}_i - \mu_i| > \delta)$ away from zero. If ε is the vector of errors, then $\hat{\mu} = Py = P(\mu + \varepsilon)$ and $\mu - \hat{\mu} = P(\mu - \mu + \varepsilon) = P\varepsilon$. Writing this row-wise, and writing $p_i{}'$ as the i-th row of P,

$$\hat{\mu}_i - \mu_i = p_i{}'\varepsilon = h_{ii}\varepsilon_i + \sum_{j \neq i}^{n} h_{ij}\varepsilon_j = w + z$$

Now, w depends only on ε_i and z on the remaining elements of ε, and so the independence assumption for elements of ε implies independence of w and z. We can therefore write

$$\begin{aligned}
\Pr(|w+z| \geq \delta) &\geq P(w \geq \delta, z \geq 0) + P(w \leq -\delta, z < 0) \\
&= P(w \geq \delta)P(z \geq 0) + P(w \leq -\delta)P(z \leq 0) \\
&\geq \min[P(w \geq \delta), P(w \leq -\delta)]
\end{aligned}$$

This implies that

$$\Pr(|\hat{\mu}_i - \mu_i| > \delta) \geq \min[\Pr(\varepsilon_i \geq \frac{\delta}{h_{ii}}), \Pr(\varepsilon_i \leq -\frac{\delta}{h_{ii}})] > 0$$

Thus, if $h_{ii} > 0$, we do not have consistency, so we must have that $h_{ii} \to 0$. Also, $\max(h_{ii}) \geq$ mean$(h_{ii}) = p/n$, hence a necessary condition is $p/n \to 0$.

What does this mean? The requirement here is the all the diagonal elements of the projection matrix, which are usually called the *leverages*, go to zero. One interpretation is that the number of parameters must not grow too fast. A second interpretation is on each of the h_{ii}.

For example, suppose we have a one-way anova problem with t groups and n_i observations in group i. The projection matrix P is given by

$$P = \text{diag}\left(J_{n_1}J_{n_1}{}'/n_1, \ldots, J_{n_t}J_{n_t}{}'/n_t\right)$$

so a typical leverage for a case in group i is $1/n_i$.

If we fix t and n_1 but allow all the other n_i to grow large, then the leverages corresponding to the first group will be fixed at $1/n_1$ while all the other leverages approach zero. Thus, all the group means except the first will be consistent. If we the n_i fixed but allow $t \to \infty$, all the leverages are fixed and non-zero, and the estimate of the mean is not consistent.

Theorem 5.2 *Let $a'\mu = a'X\beta$ be any estimable function of the mean in a general linear model, scaled so that $\| Xa \|^2 = 1$. Then var$(a'\hat{\mu}) = \sigma^2$, and $a'\hat{\mu}$ is consistent for $a'\mu$ if and only if $\max_i |s_i| \to 0$, where s_i is the i-th component of the vector $X'a$.*

Proof. Huber (1981, page 159).

Theorem 5.3 *If $\max_i h_{ii} \to 0$, then $a'\hat{\mu}$ is asymptotically normally distributed for all $a \in \Re^n$. If not, then some linear combinations of the $\hat{\mu}_i$ are not asymptotically normal.*

Proof. Huber (1981, p. 159)

This last theorem is the key to understanding the asymptotic behavior of estimates in the linear model without the assumption of normality of the errors. We get asymptotic normality as long as the leverages all go to zero as the sample size increases. If any of the leverages are large, then the asymptotic distribution of some linear combinations may not be normal.

Theorem 5.4 **Li-Duan theorem.** *Suppose that x has a multivariate distribution, and without loss of generality, assume that E$(x) = 0$ and var$(x) = I$. Suppose* *further that $E(y|x) = g(x'\beta)$ for some unknown and unspecified smooth function g. If*

$$E(a'x|b'x) = \alpha_0 + \alpha_1(b'x)$$

for all vectors a and b, or equivalently

$$E(x|Bx) = P_B x$$

for all matrices of appropriate dimension B. Then the OLS estimate of β from fitting a linear model with $\mathcal{E} = \mathcal{R}(X)$ is a consistent estimate of $c\beta$ for some constant c that is typically nonzero.

Proof. We simplify the problem by assuming that $(X'X) = I$, and so the ols estimator is just $\hat{\beta} = X'y$. This can always be attained by appropriate linear transformation of the predictors. We examine the expectation of the product xy:

$$\begin{array}{rcll} \mathrm{E}(xy) & = & \mathrm{E}[\mathrm{E}(xy|\beta'x, y)] & \ldots \textit{iterated expectations} \\ & = & \mathrm{E}[\mathrm{E}(x|\beta'x, y)y] & \\ & = & \mathrm{E}[\mathrm{E}(x|\beta'x)y] & \ldots \textit{conditional independence} \\ & = & \mathrm{E}[(P_\beta x)y] & \ldots \textit{linear predictors} \\ & = & P_\beta[\mathrm{E}(xy)] & \ldots \textit{the projection matrix is constant} \end{array}$$

We have shown that $\mathrm{E}(xy) \in \mathcal{R}(\beta)$, so it has expectation $c\beta$ for some non-zero constant c, and so $\mathrm{E}(X'y) = c\beta$.

5.2 The Normal Distribution

Most of linear model theory, for example testing and other inference, will require some specification of distributions, and, of course the normal or Gaussian distribution plays the central role here.

Definition 5.2 (Univariate Normal distribution) *The univariate normal density with respect to Lebesgue measure is*

$$f(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \tag{5.3}$$

where $\sigma > 0$ and $-\infty < y < \infty$. We usually write $y \sim N(\mu, \sigma^2)$.

Normal distributions depend only on first and second moments; all higher moments depend only on constants, μ and σ^2.

Definition 5.3 (Multivariate Normal Distribution) *A random vector y has a multivariate normal distribution if $a'y$ has a univariate normal distribution for all $a \in \Re^n$. If $E(y) = \mu$ and $Var(y) = \Sigma$, and y has a multivariate normal distribution, then*

$$a'y \sim N(a'\mu, a'\Sigma a)$$

so that μ and Σ characterize the multivariate normal and we can write $y \sim N(\mu, \Sigma)$.

Theorem 5.5 *If $y \sim N(\mu, \Sigma)$ and $A : n \times p$ and $b : p \times 1$ are fixed, then*

$$Ay + b \sim N(A\mu + b, A\Sigma A').$$

Proof. Follows from the previous construction method: $a'(Ay + b) = (a'A)y + a'b \sim N(a'A\mu + a'b, a'A\Sigma A'a)$ for all a.

5.2.1 Characteristic functions

We define characteristic functions and then give several of their properties without proof.

Definition 5.4 *The characteristic function of a random n-vector z is*

$$\varphi_z(t) = E(\exp(it'z)),\ t \in \Re^n$$

The characteristic function of the univariate normal, $y \sim$ N$(0, 1)$ is

$$\mathrm{E}(\exp(ity)) = \exp(-t^2/2) \tag{5.4}$$

Suppose α is a nonzero scalar, and b is also a scalar. Then

$$\varphi_{z+b}(t) = \exp(it'b)\varphi_z(t) \tag{5.5}$$
$$\varphi_{\alpha z}(t) = \varphi_z(\alpha t) \tag{5.6}$$

Using these two results, the characteristic function of $z = \mu + \sigma y$ where $y \sim$ N$(0, 1)$ is

$$\varphi_{\mu+\sigma y}(t) = \exp(it\mu)\varphi_y(\sigma t) = \exp(it\mu - t^2\sigma^2/2)$$

This is the characteristic function of a N(μ, σ^2) random variable. For an $m \times n$ matrix A, the characteristic function of Az is

$$\varphi_{Az}(s) = \mathrm{E}(\exp(is'Az)) = \varphi_z(A's) \tag{5.7}$$

where s is $m \times 1$. There is a one to one correspondence between characteristic functions on $\Re^n$ and distributions on $\Re^n$, so finding a characteristic function is equivalent to finding a distribution.

The random vectors w and z are *independent* if and only if their joint characteristic function factors into a product of their marginal characteristic functions:

$$\varphi_{[w,z]}\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \varphi_z(t_1)\varphi_w(t_2) \tag{5.8}$$

Suppose $z_1, \ldots, z_n$ are independent and identically distributed $\mathrm{N}(0,1)$. Then we can apply (5.8) to find the characteristic function of $z = (z_1, \ldots, z_n)$ to be

$$\begin{aligned} \varphi_{[z_1, \ldots, z_n]}(t) &= \prod_{i=1}^{n} \varphi_{[z_i]}(t_i) \\ &= \prod_{i=1}^{n} \exp(-t_i^2/2) \\ &= \exp(-t't/2) \end{aligned}$$

where $t = (t_1, \ldots, t_n)'$. To find the characteristic function of $y = \mu + \Sigma^{1/2}z$, which of course follows a $\mathrm{N}(\mu, \Sigma)$ distribution, we need only apply (5.5) and (5.7),

$$\begin{aligned} \varphi_{[y]}(t) &= \varphi_{[\mu+\Sigma^{1/2}z]}(t) \\ &= \exp(it'\mu)\varphi_{[\Sigma^{1/2}z]}(t) \\ &= \exp(it'\mu)\varphi_{[z]}(\Sigma'^{1/2}t) \\ &= \exp(it'\mu - t'\Sigma t/2) \end{aligned}$$

5.2.2 More independence

Theorem 5.6 *The n-vector z and the m-vector w are independent if and only if $a'z$ and $b'w$ are independent for all $a \in \Re^n$ and $b \in \Re^m$.*

Proof. Homework.

We turn now specifically to the multivariate normal distribution, and consider arbitrary linear combinations $a'y$ and $b'y$. We know that, for general distributions, the notion of uncorrelated is not equivalent to the notion of independent. For the normal case, however, we can get a stronger result.

Theorem 5.7 (Independence of linear combinations) *If $y \sim N(\mu, \Sigma)$, then $a'y$ and $b'y$ are independent if and only if they are uncorrelated.*

Proof. $a'y$ and $b'y$ are independent if and only if

$$\varphi_{[a'y,b'y]}\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \varphi_{a'y}(t_1)\varphi_{b'y}(t_2) \tag{5.9}$$

for all t_1, t_2. The three characteristic functions are easily computed to be

$$\begin{aligned}
\varphi_{[a'y,b'y]}\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} &= \exp(i(t_1a' + t_2b')\mu - .5(t_1a + t_2b)'\Sigma(t_1a + t_2b)) && (5.10)\\
\varphi_{[a'y]}(t_1) &= \exp(it_1a'\mu - .5t_1^2a'\Sigma a) && (5.11)\\
\varphi_{[b'y]}(t_1) &= \exp(it_1b'\mu - .5t_1^2b'\Sigma b) && (5.12)
\end{aligned}$$

The left-hand side of (5.9) is given by (5.10) and the right-hand side is given by the product of (5.11) and (5.12). After a little algebra, we will see that these two are equal if and only if

$$t_1t_2a'\Sigma b = 0$$

for all t_1 and t_2. Thus $a'y$ and $b'y$ are independent if and only if $a'\Sigma b = 0$, or equivalently, if and only if they are uncorrelated.

Theorem 5.8 (Independence of linear forms) $y \sim N(\mu, \Sigma)$. *Then Ay and By are independent if and only if $A\Sigma B' = 0$. If $\Sigma = I$, then the condition is $AB' = 0$.*

Proof. This is a generalization of the Theorem 5.7. Since y is normally distributed, so are Ay and By. Then Ay and By are independent if and only if $a'Ay$ and $b'By$ are independent for all a and b, and by Theorem 5.7, this holds if and only if $a'A\Sigma B'b = 0$ for all a and b so $A\Sigma B' = 0$.

Theorem 5.9 (Independence of orthogonal projections) *Suppose $y \sim N(\mu, \sigma^2 I)$, and $P_1, \ldots, P_k$ are pairwise orthogonal projections onto $M_1, \ldots, M_k$ respectively, with $\sum P_i = I$ or equivalently, $\sum \mathcal{R}(P_i) = \Re^n$. Then:*

1. *$P_1y, \ldots, P_ky$ are mutually independent random vectors.*
2. *$P_iy \sim N(P_i\mu, \sigma^2 P_i)$.*

Proof. For the first result, $P_iP_j = 0$ implies pairwise uncorrelated and hence independence. Mutual independence follows from factorization of the characteristic function as was done to find the characteristic function of the standard multivariate normal, from (5.7).

For the second part of the theorem, we can use a characteristic function argument, as follows:

$$\varphi_{[P_i y]}(t) = \varphi_y(P_i t) = \exp(it'P_i\mu - t'P_i t/2)$$

which we recognize as the unique characteristic function of a $\mathrm{N}(P_i\mu, \sigma^2 P_i)$ random variable.

We can extend this to looking separated at each dimension of $\Re^n$ separately. Suppose $\{x_1, \ldots, x_n\}$ is an orthonormal basis for $\Re^n$. Then

$$y = \sum_{i=1}^{n}(x_i, y)x_i = \sum_{i=1}^{n} P_i y$$

is the sum of the projections onto each basis vector. By application of Theorem 5.9,

$$(x_i, y) = x_i{}'y \sim N(x_i{}'\mu, \sigma^2 x_i{}'P_i x_i)$$

Now if $\mathrm{Var}(y) = \sigma^2 I$, then $P_i = x_i x_i{}'$, and $\sigma^2 x_i{}'P_i x_i = \sigma^2$. Thus each coordinate of y relative to any orthonormal basis is normal with appropriate mean and variance σ^2. These coordinates are independent if the original ys are independent.

Similarly, if Γ is an orthogonal matrix, $\mathrm{Cov}(y) = \sigma^2 I$, then $\Gamma y \sim N(\Gamma\mu, \sigma^2 I)$.

5.2.3 Density of the Multivariate Normal Distribution

Suppose $z_1, \ldots, z_n$ are iid $\mathrm{N}(0, 1)$. Then the density of

$$z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \sim \mathrm{N}(0, I)$$

is

$$f(z) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-z_i^2/2) = \frac{1}{(2\pi)^{n/2}} \exp(-z'z/2)$$

To find the density of $Y \sim N(\mu, \Sigma)$, for $\Sigma > 0$, we use a standard change-of-variables argument (e.g., Casella and Berger, 1990, Sec. 4.3), and

$$g(y) = \frac{1}{(2\pi)^{n/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right)$$

5.3 Chi-squared distributions

The chi-squared is the most important of the many distributions that are derived from the normal distribution. These distributions play a fundamental role in linear models since the distribution of quadratic forms and in particular lengths of projections will have distributions related to the chi-squared.

Definition 5.5 (Central chi-squared random variable) *Suppose that* $z_1, \ldots, z_m$ *are independent and identically distributed (iid)* $N(0,1)$. *Then* $X = \sum_{i=1}^m z_i^2 \sim \chi^2(m)$, *a central chi-squared random variable with* m *degrees of freedom (df).*

If $X \sim \chi^2(m)$, then X has density

$$f(x) = \frac{x^{(m-2)/2} \exp(-x/2)}{2^{m/2}\Gamma(m/2)}$$

for $x \geq 0$, and zero otherwise. The characteristic function is

$$\varphi_x(t) = (1 - 2it)^{-n/2}$$

We now add the assumption of normality to the linear model to get distributions of functions of quadratic forms.

Theorem 5.10 *Consider the standard coordinate free normal linear model* $y \sim N(\mu, \sigma^2 I)$, $\mu \in \mathcal{E}$ *of dimension* p. *Then if* P *is the orthogonal projection on* $\mathcal{E}$,

$$\frac{\| P(y-\mu) \|^2}{\sigma^2} \sim \chi^2(p)$$

Proof. Let $x_1, \ldots, x_p$ be an orthonormal basis for $\mathcal{E}$. Then

$$P(y-\mu) = \sum_{i=1}^p (x_i, y-\mu)x_i$$

and

$$\| P(y-\mu) \|^2 = \sum_{i=1}^p (x_i, y-\mu)^2$$

But $(x_i, y-\mu) \sim N(0, \sigma^2)$, and are independent for each i. Thus $(x_i, y-\mu)/\sigma^2 \sim N(0,1)$, and so the result follows by the definition of the central chi-squared.

Theorem 5.11 *Suppose that $\mathcal{E}_1, \ldots, \mathcal{E}_k$ is an orthogonal partition of the p-dimensional subspace $\mathcal{E} \subset \Re^n$, and P_i are the associated orthogonal projection operators, so $\mathcal{E} = \sum \mathcal{E}_i$, $P_{\mathcal{E}} = \sum P_i$, and $\sum \dim(\mathcal{E}_i) = p$. Then:*

$$\frac{\| P(y-\mu) \|^2}{\sigma^2} = \sum_{i=1}^{k} \frac{\| P_i(y-\mu) \|^2}{\sigma^2} \tag{5.13}$$

where the left side of (5.13) is distributed as a $\chi^2(p)$, and each term on the right side is distributed as an independent $\chi^2(\dim(P_i))$.

Theorem 5.12 *In the standard coordinate free normal linear model $y \sim N(\mu, \sigma^2 I)$, $\mu \in \mathcal{E}$ of dimension p, if P is the orthogonal projection on $\mathcal{E}$, then*

$$\frac{\| (I-P)y \|^2}{\sigma^2} \sim \chi^2_{n-p}$$

Generalized least squares In the linear model $y \sim \mathrm{N}(\mu, \sigma^2\Sigma)$ with $\Sigma = \Sigma' > 0$ known, Theorems 5.11–5.12 still hold, if we replace the inner product $(a, b) = a'b$ with the inner product $(a, b)_{\Sigma^{-1}} = (a, \Sigma^{-1}b)$ and replace the norm $\| a \|^2 = a'a$ by the norm $\| a \|^2_{\Sigma^{-1}} = a'\Sigma^{-1}a$.

5.3.1 Non-central χ^2 Distribution

Definition 5.6 *Let $z_1, \ldots, z_m$ be independent $N(\mu_i, 1)$ random variables. Define $X = \sum_{i=1}^m, z_i^2$ to be a non-central chi-squared random variable with degrees of freedom m, and non-centrality parameter $\delta^2 = \sum_{i=1}^m, \mu_i^2 = \mu'\mu$.*

The density of X can be written, for $x > 0$ as:

$$f(x) = \sum_{k=0}^{\infty} \left[\frac{\exp(-\delta^2/2)(\delta^2/2)^k}{k!} \right] \frac{x^{(2k+m-2)/2} \exp(x/2)}{2^{(2k+m)/2}\Gamma[(2k+m)/2)]}$$

This is a Poisson mixture of central χ^2s with degrees of freedom $2k+m$ and with $k \sim \mathrm{Po}(\delta^2/2)$. Sometimes, $\lambda = \delta^2/2$, the mean of the Poisson, is used for the non-centrality parameter.

Suppose that $X_i \sim \chi^2(m_i, \delta_i^2), i = 1, 2$. The mean and variance of X_1 are given by

$$\begin{aligned} \mathrm{E}(X_1) &= m_1 + \delta_1^2 &(5.14)\\ \mathrm{Var}(X_1) &= 2(m_1 + 2\delta_1^2) &(5.15) \end{aligned}$$

If X_1 and X_2 are independent, then $X_1 + X_2 \sim \chi^2(m_1 + m_2, \delta_1^2 + \delta_2^2)$. The characteristic function of $X = X_1$ is given by

$$\begin{aligned} \varphi_x(t) &= \sum_{k=0}^{\infty} \frac{\exp(-\delta^2/2)(\delta^2/2)^k}{k!} \frac{1}{(1-2it)^{k+m/2}} \\ &= \frac{\exp(-\delta^2/2)}{(1-2it)^{m/2}} \sum_{k=0}^{\infty} \frac{(\delta^2/2)[1/(1-2it)]^k}{k!} \end{aligned}$$

The infinite sum in this last expression is just the sum of terms for the expectation of a Poisson variate t, where $t \sim Po(\delta^2/2)$, so $\mathrm{E}[1/(1-2it)] = \exp(\delta^2/[2(1-2it)])$. Substituting, we get

$$\begin{aligned} \varphi_x(t) &= \frac{1}{(1-2it)^{m/2}} \exp\left(\frac{\delta^2}{2}\left[\frac{1}{1-2it} - 1\right]\right) \\ &= \frac{1}{(1-2it)^{m/2}} \exp\left(\frac{\delta^2 it}{1-2it}\right) \end{aligned}$$

Return to the general coordinate free normal linear model, $y \sim N(\mu, \sigma^2\Sigma)$ with $\Sigma = \Sigma' > 0$ known. Use the norm $(a,b)_{\Sigma^{-1}} = (a, \Sigma^{-1}b)$ and norm $\| a \|^2_{\Sigma^{-1}} = a'\Sigma^{-1}a$ and suppose that $P_1, \ldots, P_k$ mutually orthogonal projections and $y = \sum P_i y$. Then $P_i y$, $i = 1, \ldots, k$ are independent:

$$\| y \|^2 = \| P_1 y \|^2 + \cdots + \| P_k y \|^2$$

and

$$\frac{\| P_i y \|^2}{\sigma^2} \sim \chi^2(\dim(P_i), \delta_i^2)$$

and $\delta_i^2 = \| P_i \mu \|^2 / \sigma^2$.

5.4 The distribution of quadratic forms

Suppose y is a random vector of length n, and A is an $n \times n$ matrix. A *quadratic form* in y is the quantity $y'Ay$. We will assume A is symmetric, but this is irrelevant since $.5(A + A')$ is always symmetric, and $y'Ay = .5y'(A + A')y$. Quantities such as these are fundamental in linear models, and we would like to know the distribution of $y'Ay$. The following theorem characterizes matrices A for which $y'Ay$ has a chi-squared distribution.

Theorem 5.13 *Let A denote a matrix of rank p and let $z \sim N(0, I)$. Then $z'Az \sim \chi^2(p)$ if and only if A is an orthogonal projection on a space of rank p.*

Proof. Assume A is an orthogonal projection. Then the spectral decomposition of A is

$$A = \Gamma \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \Gamma'$$

so

$$z'Az = z'\Gamma \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \Gamma' z = z'\Gamma_1 {\Gamma_1}' z$$

where Γ_1 is the first p columns of Γ. Since ${\Gamma_1}'z \sim N(0, \sigma^2 {\Gamma_1}'\Gamma_1)$, and ${\Gamma_1}'\Gamma_1 = I$, so by definition $z'Az \sim \chi^2(p)$.

Next, assume $z'Az \sim \chi^2(p)$. Again write A in spectral form, but now we do not assume that A is a projection, so the decomposition is

$$A = \Gamma \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} \Gamma'$$

and the diagonals of Λ positive but need not all be one. However, by assumption

$$\sum_{i=1}^{p} \lambda_i z_i^2 \sim \chi^2(p) \tag{5.16}$$

and each $z_i \sim N(0, 1)$. We need to show that (5.16) implies that each of the λ_i must be equal to 1. This can be done by calculating the characteristic function of $\sum \lambda_i z_i^2$ and show that this is the characteristic function of $\chi^2(p)$ if and only if $\lambda_i = 1$.

A general quadratic form is given by

$$Q = (y - \mu)'G(y - \mu) \tag{5.17}$$

where $y \sim N(\mu, \Sigma)$, G is an arbitrary real symmetric matrix, and Σ of full rank. Suppose that B is a square root of Σ, so $\Sigma = BB'$. We can write

$$Q = (y - \mu)'B^{-1}BGB'B'^{-1}(y - \mu) = z^{*\prime}BGB'z^*$$

where $z^* \sim \mathrm{N}(0, I)$. Writing the spectral decomposition of BGB',

$$Q = z^{*\prime}\Gamma \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} \Gamma' z^* = z^{**\prime} \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} z^{**}$$

where $z^{**} = \Gamma z^*$ is also $\mathrm{N}(0, I)$. Finally, we get

$$Q = \sum_{i=1}^{p} \lambda_i^2 z^{**2} = \sum_{i=1}^{p} \lambda_i^2 \chi^2(1)$$

since each of the z_i^{**2} have independent $\chi^2(1)$ distributions, and the λ_i are the eigenvalues of BGB'. This shows that the distribution of a general quadratic form of normal random variables (5.17) is distributed as a sum of $\chi^2(1)$ random variables. Johnson and Kotz (1972, *Distributions in Statistics: Continuous Univariate Distributions, Vol. II*, p. 150–153) provide discussion, and Bowman and Azzalini (1997, *Applied Smoothing Techniques for Data Analysis*, p. 87) discusses computing percentage points for the ratio of two such quadratic forms.

Finally, we give a version of Cochran's Theorem:

Theorem 5.14 *Let $A_1, \ldots, A_m$ by $n \times n$ symmetric matrices, and $A = \sum A_j$, with rank$(A_j) = n_j$. Consider the following four statements:*

- *A_j is an orthogonal projection for all j.*
- *A is an orthogonal projection (possibly, $A = \mathbf{I}$).*
- *$A_j A_k = 0$ for all (j, k).*
- *$\sum n_j = n$.*

If any two of these conditions hold, then all four hold.

The application of this theorem to linear models is that if Theorem 5.14 holds, and $y \sim N(\mu, \sigma^2 I)$, then: $Q_j = y' A_j y \sim \chi^2(n_j, \mu' A_j \mu / \sigma^2)$ and $Q_1, \ldots Q_m$ are independent.

Cochran's theorem is a standard result that is the basis of the analysis of variance. If $y' A_j y$ has a Chi-squared distribution, then A_j must be a projection matrix but this theorem says more. If we can write the total sum of squares as a sum of sum of squares components, and if the degrees of freedom add up, then the A_j must be projections, they are orthogonal to each other, and they jointly span $\Re^n$!

5.5 The Central and Non-Central F-distribution

Suppose that $X_1 \sim \chi^2(n_1, \delta_1^2)$ and $X_2 \sim \chi^2(n_2, \delta_2^2)$, with X_1 and X_2 independent. Define the ratio:

$$F = \frac{X_1 / n_1}{X_2 / n_2} \tag{5.18}$$

The distribution of F is called the *doubly non-central F-distribution* with degrees of freedom (n_1, n_2) and non-centrality parameters (δ_1^2, δ_2^2). If in (5.18) $\delta_2^2 = 0$, we write $F(n_1, n_2; \delta_1^2)$, and refer to this as the *non-central* F-distribution. If $\delta_1^2 = \delta_2^2 = 0$ in (5.18), we write $F(n_1, n_2)$, and we call (5.18) a *central* F distribution. The doubly non-central F arises only in nonstandard situations that we may encounter later. Standard applications use the central F and the non-central F.

Suppose that $f \sim F(n_1, n_2, \delta^2)$, then for $n_2 > 2$,

$$\mathrm{E}(f) = \frac{n_2(n_1 + \delta^2)}{n_1(n_2 - 2)}$$

and, for $n_2 > 4$, the variance is

$$Var(f) = 2\left(\frac{n_2}{n_1}\right)^2 \frac{(n_1 + \delta)^2 + (n_1 + 2\delta^2)(n_2 - 2)}{(n_2 - 2)^2(n_2 - 4)}$$

Existence of the third central moment requires $n_2 > 6$. If $f \sim F(n_1, n_2)$, so f has a central distribution, the density of f can be written, for $f > 0$, as:

$$h(f) = \frac{n_1^{n_1/2} n_2^{n_2/2} f^{(n_1/2-1)}}{B(n_1/2, n_2/2)(n_2 + n_1 f)^{(n_1+n_2)/2}}$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ is the beta function. The density of the non-central F can be written as a Poisson mixture of densities that are proportional to those of central fs: Write $f_{a,b} \sim F(a, b)$; $g_{a,b} = (a/b)f_{a,b} = \chi^2(a)/\chi^2(b)$; and $f \sim F(n_1, n_2, \delta^2)$ and $g = \chi^2(n_1, \delta^2)/\chi^2(n_2)$, so $f = n_2 g/n_1$. Then:

$$F(g) = \sum_{k=0}^{\infty} \frac{\exp(-\delta^2/2)(\delta^2/2)k}{k!} G_{g_{n_1+2k,n_2}}(f)$$

The non-central F is not a Poisson mixture of central Fs.

5.6 Student's t distribution

Closely related to the F-distribution is Student's t-distribution.

Definition 5.7 (Student's t-distribution) *Suppose that $z \sim N(\mu, \sigma^2)$, and $s^2 \sim \sigma^2\chi^2(d)$ such that z and s^2 are independent. Then the ratio*

$$t = \frac{z}{s}$$

is distributed as a non-central Student's t *distribution with* d *degrees of freedom, and non-centrality parameter* $\delta = \mu^2$. *The* central Student's t*-distribution has* $\mu = 0$.

The density function of the central t distribution is

$$f(t) = \left(d^{1/2} B(1/2, d/2)(1 + t^2/d)^{(d+1)/2}\right)^{-1}$$

where $B(a, b)$ is the Beta function. The density is symmetric about zero, has mean zero, and variance $d/(d-2)$ for $d > 2$. The case $d = 1$ corresponds to the Cauchy distribution. From the definition, we see that $t^2 \sim F(1, d)$, a central F with 1 and d df.

In the normal linear model $y \sim \mathrm{N}(\mu, \sigma^2 I)$ with $\mu \in \mathcal{E}$, a p-dimensional subspace, an unbiased estimate of σ^2 is given by $\hat{\sigma}^2 = y'(I-P)y/(n-p)$. Let $a'\hat{\mu}$ be the estimator of an estimable function, and so $a'\hat{\mu} \sim \mathrm{N}(a'\mu, \sigma^2 a'Pa)$, independent of $\hat{\sigma}^2$. Hence the ratio

$$t = \frac{a'\hat{\mu} - a'\mu}{\hat{\sigma}\sqrt{a'Pa}}$$

is distributed as Student's t with $n - p$ df.