Statistical Inference Theory

Rebecca Sela

September 28, 2005

1 Preliminaries

Statistical inference is based on a model given by $\{P_\theta : \theta \in \Theta\}$. We then draw observations, $x_1, ..., x_n$ of the random vector X which are (usually) independent and identically distributed from the distribution P_{θ_0}, where θ_0 is the true parameter value. θ_0 can never be known; statistical inference attempts to estimate it as precisely as possible.

Definition If $\Theta \subset R^k$, for finite k, then the model is called *parametric*. Otherwise, the model is *non-parametric* (or *semi-parametric*).

It is always possible for the underlying set of models to be wrong. For example, the distribution may be different, or the observations may not be independent. This is not the problem of statistical inference; this is model checking.

Definition Suppose we observe x from the model $\{P_\theta : \theta \in \Theta\}$. The *likelihood function* is given by:

$$lik(\theta; x) = P_\theta(x)$$

The likelihood function is a function of θ, while the probability density function is a function of x.

2 Criteria for choosing an estimator

Definition An *estimator* is a statistic (that is, a mapping from random variables to other random variables) that attempts to get close to a function of a parameter. An *estimate* is the value that the estimator takes for the particular data observed. To estimate $g(\theta)$, we have an estimator $\hat{g}(X)$ and an estimate $\hat{g}(x)$.

An estimator cannot be correct with probability one except in very unusual situations (such as if the data partition the space). However, for very large sample sizes this is approximately possible.

Definition An estimator is (strongly) *consistent* if

$$\lim_{n\to\infty} P_\theta\left(\hat{g}(X_n) = g(\theta)\right) = 1$$

In finite samples, we hope that the estimator is close to the true value of the parameter.

Definition An estimator $\hat{g}$ is *unbiased* if $E_\theta(\hat{g}(X)) = g(\theta)$ for all $\theta \in \Theta$.

Definition The *mean square error* of an estimator $\hat{g}$ of $g(\theta)$ is $E_\theta\left((\hat{g}(X) - g(\theta))^2\right)$.

Proposition 2.1 *The mean square error of an estimator is the sum of its variance and its squared bias. That is,*

$$E_\theta\left((\hat{g}(X) - g(\theta))^2\right) = Var(\hat{g}(X)) + (E(\hat{g}(X)) - g(\theta))^2$$

3 Sufficiency

It is not always necessary to know the values and order of each observation in order to do inference. We wish to reduce the data to a smaller number of statistics that we may use instead. Statistics that summarize the data without losing any information about the parameters are called sufficient (and vary with the model being considered). This creates a partition of the space of outcomes, in which samples are equivalent if they have the same sufficient statistics.

Definition Let Ω be the sample space of X. Let $\{P_\theta : \theta \in \Theta\}$ be a family of models. Let Π be a partition of Ω. Π is a *sufficient partition* with respect to $\theta \in \Theta$ if $P_\theta(X \mid A \in \Pi)$ does not depend on θ for all $A \in \Pi$.

Definition Let $T : \Omega \to R^k$ be a function which is constant on each sufficient partition and takes a different value for different partitions (that is, each $T^{-1}(a)$ corresponds to exactly one (possibly empty) set of the sufficient partition). Then, T is a *sufficient statistic*, and $P_\theta(X \mid T(X) = t)$ does not depend on θ for all $t \in R^k$.

Notice that any one-to-one function of a sufficient statistic is also a sufficient statistic, so that sufficient statistics are not unique.

Theorem 3.1 (Factorization Theorem) *Given a family of densities, $\{P_\theta : \theta \in \Theta\}$, $T(X)$ is a sufficient statistic if and only if $P_\theta(X) = P_\theta(T)h(X)$, where $h(X)$ does not depend on θ.*

Definition A *minimal sufficient partition* is the coarsest partition which is still sufficient. (This is the maximal data reduction possible.) A *minimal sufficient statistic* is a statistic that takes a distinct value on each minimal sufficient partition.

If there is a minimal sufficient partition, then the sets in any other sufficient partition must be subsets of the minimal partition.

Theorem 3.2 *Given* $\{P_\theta : \theta \in \Theta\}$*, define a partition made up of the sets* $\overline{Y} = \{X : P_\theta(X)/P_\theta(Y)$ *does not depend on* $\theta\}$ *for each* Y*. This partition is minimal sufficient.*

Proof Suppose $\Pi = \{\{X : P_\theta(X)/P_\theta(Y)$ does not depend on $\theta\} : Y \in \Omega\}$. This is a partition. Furthermore, this partition is sufficient, since $P_\theta(X \mid X \in \overline{X}) = P_\theta(X)/\sum_{Y \in \overline{X}} P_\theta(Y)$ is the reciprocal of the sum of $P_\theta(Y)/P_\theta(X)$, none of which depend on θ, so the probability of a certain sample given the partition does not depend on θ.

Let T be any sufficient statistic. Suppose $T(Y) = T(X)$ because they are in the same partition, and, by the Factorization Theorem:

$$\frac{P_\theta(X)}{P_\theta(Y)} = \frac{P_\theta(T(X))h(X)}{P_\theta(T(Y))h(Y)} = \frac{h(X)}{h(Y)}$$

This does not depend on θ. Thus, Y and X must be in the same set of Π, and the sets of the partition defined by any sufficient statistic are subsets of the partition above. ∎

Theorem 3.3 (Rao-Blackwell Theorem) *Suppose* $\tilde{g}(X)$ *is an unbiased estimator of* $g(\theta)$*. Suppose* $T(X)$ *is a sufficient statistic for the family of distributions. Then,* $\hat{g}(X) = E(\tilde{g}(X) \mid T(X))$ *is an unbiased estimator of* $g(\theta)$ *with at most the variance of* $\tilde{g}(X)$*.*

Proof $\hat{g}(X) = E(\tilde{g}(X) \mid T(X))$ is an estimator because $E(\tilde{g}(X) \mid T(X))$ does not depend on θ since $P(\tilde{g}(X) \mid T(X))$ does not depend on θ. By the law of iterated expectations,

$$\begin{aligned} E\left(E_\theta(\tilde{g}(X) \mid T(X))\right) &= E(\tilde{g}(X)) \\ &= g(\theta) \end{aligned}$$

For any random variables X and Y,

$$Var(X) = E(Var(X \mid Y)) + Var(E(X \mid Y)) \geq Var(E(X \mid Y))$$

so,

$$Var(\tilde{g}(X)) \geq Var(E(\tilde{g}(X) \mid T)) = Var(\hat{g}(X))$$

∎

In general, conditioning any estimator on a sufficient statistic tends to reduce the mean square error.

4 Completeness

Definition Let $\{P_\theta(T) : \theta \in \Theta\}$ be a family of distributions of a sufficient statistic T. This family is *complete* if $E_\theta(f(T)) = 0$ for all θ implies that $f(t) = 0$ almost everywhere. The family is *bounded complete* if $E_\theta(f(T)) = 0$ for a bounded f implies that $f(t) = 0$ almost everywhere.

Theorem 4.1 *If $\{P_\theta(T) : \theta \in \Theta\}$ is complete, then T is a minimal sufficient statistic for the original family.*

Proof *(Sketch.)* Suppose t_1 and t_2 are in a single partition element, A, defined by a sufficient partition, t. Then the probabilities $f_\theta(t_1 \mid A) = p$ and $f_\theta(A - t_1 \mid A) = 1 - p$ do not depend on θ. Set $h(t_1) = -1/p$ and $h(A - t_1) = 1/(1 - p)$. Then $E_\theta(h|A) = 0$. This is a contradiction. ∎

Note that minimal sufficient does not imply complete. (For example, combining two experiments that measure the same parameter in different ways may make a family incomplete.)

Theorem 4.2 *If T is a complete sufficient statistic, then any unbiased estimator $\tilde{g}(X)$ of $g(\theta)$ can be used to find the minimum variance unbiased estimator by conditioning on T. That is, the minimum variance unbiased estimator is $\hat{g} = E(\tilde{g} \mid T)$. The minimum variance estimator is unique (except for a set of measure 0).*

Definition If $P_\theta(X) = C(\theta) \exp(\sum_{i=1}^{k} Q_i(\theta)T_i(X))h(X)$, for known functions C, Q_i, T_i, and h, then P_θ is in the *exponential family.*

By the factorization theorem, we see that $T = (T_1, ..., T_k)$ is a sufficient statistic for θ.

Theorem 4.3 *Suppose we have an exponential family of distributions, $P_\theta(X) = C(\theta) \exp(\sum_{i=1}^{k} Q_i(\theta)T_i(X))h(X)$. Define $\phi_i = Q_i(\theta)$, so that $P_\phi(t) = d(\phi) \exp(\sum_{i=1}^{k} \phi_i t_i)$ (this is possible in most cases). If the transformed Θ contains a k-ball in R^k, then T is complete sufficient for ϕ.*

So, in certain cases we may find the minimum variance unbiased estimator in this way:

- Find a sufficient statistic, T, using the factorization theorem or equivalence classes.
- Prove that the family with this sufficient statistic is complete, often by showing that the statistic belongs to the exponential family and that the parameter space contains a k-ball.
- Find any unbiased estimator, $\tilde{g}(X)$. (This may depend on only one or two of the observations.)
- Compute $E(\tilde{g}(X) \mid T)$. (This will require finding $P(X \mid T)$.)

5 Efficiency

Definition $I_n(\theta) = E((\frac{\partial \ln P_\theta(X_1,\ldots,X_n)}{\partial\theta})^2)$ is called *Fisher's Information Criterion*, or the information contained in X. For a single observation, we have $i(\theta) = E_\theta((\frac{\partial \ln P_\theta(X)}{\partial\theta})^2)$, and for independent observations, $I_n(\theta) = n \cdot i(\theta)$.

Proposition 5.1 *Under certain regularity conditions,* $E(\frac{\partial}{\partial\theta}\ln P_\theta(x)) = 0$.

Proof

$$\begin{aligned} E(\frac{\partial}{\partial\theta}\ln P_\theta(x)) &= \int \frac{\partial \ln P_\theta(x)}{\partial\theta} P_\theta(X)dx \\ &= \int \frac{\partial P_\theta(X)}{\partial\theta}\cdot\frac{1}{P_\theta(x)}\cdot P_\theta(x)dx \\ &= \int \frac{\partial}{\partial\theta}P_\theta(x)dx \\ &= \frac{\partial}{\partial\theta}\int P_\theta(x)dx \\ &= \frac{\partial}{\partial\theta}(1) \\ &= 0 \end{aligned}$$

Proposition 5.2 *Under regularity conditions,* $E_\theta((\frac{\partial}{\partial\theta}\ln P_\theta(x))^2) = -E_\theta(\frac{\partial^2}{\partial\theta^2}\ln P_\theta(x))$.

Proof It is sufficient to prove this for a sample size of one, because of independence.

$$\begin{aligned} -\frac{\partial^2}{\partial\theta^2}\ln P_\theta(x) &= -\frac{\partial}{\partial\theta}(\frac{\frac{\partial}{\partial\theta}P_\theta}{P_\theta}) \\ &= -(\frac{(\frac{\partial^2}{\partial\theta^2}P_\theta)\cdot P_\theta}{P_\theta^2} - \frac{(\frac{\partial}{\partial\theta}P_\theta)^2}{P_\theta^2}) \\ &= -\frac{\frac{\partial^2}{\partial\theta^2}P_\theta}{P_\theta} + (\frac{\frac{\partial}{\partial\theta}P_\theta}{P_\theta})^2 \end{aligned}$$

Taking expectations of both sides, we have:

$$E(-\frac{\partial^2}{\partial\theta^2}\ln P_\theta(x)) = -E_\theta(\frac{\frac{\partial^2}{\partial\theta^2}P_\theta}{P_\theta}) + i(\theta)$$

The first term on the right hand side is zero (once again writing the expectation as an integral and interchanging the order of integration and differentiation). Thus, $E(-\frac{\partial^2}{\partial\theta^2}\ln P_\theta(x)) = i(\theta)$.

Theorem 5.3 Cramer-Rao Lower Bound *Under certain regularity conditions, for any unbiased estimator,* $\hat{\theta}(x)$ *of* θ, $Var(\hat{\theta}(x)) \geq 1/I_n(\theta)$.

Proof Since $\hat{\theta}(X)$ is unbiased, $E_\theta(\hat{\theta}(x)) = \theta$. We take the derivative of this with respect to θ:

$$\begin{aligned}
1 &= \frac{\partial}{\partial\theta}\theta \\
&= \frac{\partial}{\partial\theta}E_\theta(\hat{\theta}(x)) \\
&= \frac{\partial}{\partial\theta}\int \hat{\theta}(x)P_\theta(x)dx \\
&= \int \hat{\theta}(x)\frac{\partial}{\partial\theta}P_\theta(x)dx \\
&= \int \hat{\theta}(x)\left(\frac{\partial}{\partial\theta}\ln P_\theta(x)\right)P_\theta(x)dx \\
&= E_\theta\left(\hat{\theta}(x)\cdot\frac{\partial}{\partial\theta}\ln P_\theta(x)\right) \\
&= Cov\left(\hat{\theta}(x),\frac{\partial}{\partial\theta}\ln P_\theta(x)\right) \\
&\leq Var(\hat{\theta}(x))Var\left(\frac{\partial}{\partial\theta}\ln P_\theta(x)\right) \\
&= Var(\hat{\theta}(x))I_n(\theta)
\end{aligned}$$

Thus, $Var(\hat{\theta}(x)) \geq 1/I_n(\theta)$, and Fisher's information is a lower bound on the variance of an unbiased estimator.

It may not be possible to achieve the Cramer-Rao lower bound, and more than one estimator may achieve this lower bound.

Theorem 5.4 *Suppose $\hat{g}(x)$ is an unbiased estimator of $g(\theta)$. Then*

$$Var(\hat{g}(x)) \geq \frac{(g'(\theta)^2)}{I_n(\theta)}$$

Theorem 5.5 *Suppose $\hat{\theta}(x)$ is an biased estimator of θ, with $E(\hat{\theta}(x)) = \theta + b(\theta)$. Then*

$$Var(\hat{\theta}(x)) \geq \frac{(1+\frac{\partial}{\partial\theta}b(\theta))^2}{I_n(\theta)}$$

Definition The *efficiency* of an unbiased estimator, $\hat{g}(x)$, with a lower bound of $b(\theta)$ on the variance is:

$$eff(\hat{g}(x)) = \frac{b(\theta)}{Var(\hat{g}(x))}$$

An estimator is *fully efficient* if its efficiency is 1.

Theorem 5.6 *Under the same regularity conditions, an unbiased estimator, $\hat{\theta}(x)$ of θ is fully efficient if and only if*

$$\frac{\partial}{\partial\theta}\ln P_\theta(x) = I_n(\theta)(\hat{\theta}(x)-\theta)$$

Proof Let $v(x) = \frac{\partial}{\partial\theta}\ln P_\theta(x)$. Then,

$$1 = Cov(v(x),\hat{\theta}(x)) \leq Var(\hat{\theta}(x))I_n(\theta)$$

If $\hat{\theta}(x)$ is fully efficient, then the inequality above becomes an equality, and the correlation between the two variables must be one. Thus, there is a linear relationship:

$$\phi(\theta)(\hat{\theta}(x)-\theta) = v(x) - E_\theta(v(x))$$

Recall that $E_\theta(v(x)) = 0$. Also,

$$\begin{aligned} I_n(\hat{\theta}) &= Var(\hat{\theta}) \\ &= Var(v(x))/(\phi(\theta))^2 \end{aligned}$$

Thus, $\phi(\theta) = I_n(\theta)$. ∎

This gives a test for whether a fully efficient unbiased estimator exists; look at the form of $\frac{\partial}{\partial\theta}\ln P_\theta(x)$.

In the multivariate case, we have a parameter $\theta = (\theta_1,...,\theta_k)$, and an estimator is unbiased if each coordinate is unbiased. The covariance matrix of an unbiased estimator is given by:

$$Var_\theta(\hat{\theta}) = E_\theta((\hat{\theta}-\theta)(\hat{\theta}-\theta)^T)$$

A function $B(\theta)$ is a lower bound on the variance if $Var(\hat{\theta}) - B(\theta)$ is positive semi-definite. The Fisher Information Matrix, $I_n(\theta)$ is a $k\times k$ matrix with the component in the (i,j) position equal to $E_\theta(\frac{\partial\ln P_\theta(x)}{\partial\theta_i}\cdot\frac{\partial\ln P_\theta(x)}{\partial\theta_j})$ (this also equals $-E_\theta(\frac{\partial^2\ln P_\theta(x)}{\partial\theta_i\partial\theta_j})$). Note that the diagonal elements are the Fisher information for the individual parameters. $I_n(\theta)^{-1}$ is a lower bound on the variance for any unbiased estimator.

Theorem 5.7 *If $\hat{\theta}$ is an unbiased estimator for θ, then $\sum_{i=1}^k a_i\hat{\theta}_i(x)$ is an unbiased estimator for $\sum_{i=1}^k a_i\theta_i$, and a lower bound on the variance of this estimator is $a^T I_n(\theta)^{-1}a$.*

Definition If an estimator has a variance lower than the Cramer-Rao lower bound, then it is called *super-efficient*.

Super-efficient estimators are not regular. Either they do not obey the regularity conditions assumed in the theorem, or they do not behave the same way on all of the real line, so that they are superefficient if the true value of θ is at one of a countable number of points and just efficient everywhere else.

6 Method of Least Squares

Suppose we have a multivariate model $X = A\beta + \epsilon$, where $E(\epsilon) = 0$, but we make no assumption about the specific distribution of ϵ. Note that X and ϵ are n-dimensional vectors, A is an $n \times k$ matrix, and β is a k-dimensional vector of parameters to be estimated. (In terms of a single observations, x_i, this is written as $x_i = a_{1i}\beta_1 + ... + a_{ik}\beta_k + \epsilon_i$ for $i = 1, ..., n$.) In this model, the parameters β are of interest, while other parameters like the distribution of ϵ are nuisance parameters. For this reason, we want an estimator that makes no additional assumptions about the distribution of ϵ. We do assume, though, that $E(X) = A\beta$.

Definition The *least squares estimator* is an estimator chosen to minimize the sum of squared estimated errors (residuals) in the model. In the case above, $\hat{\beta}$ is chosen to minimize $(X - A\beta)^T(X - A\beta)$.

If we take the derivative of $(X - A\beta)^T(X - A\beta)$ with respect to β and set it equal to 0, we find the normal equation:

$$A^T A\beta = A^T X$$

Our estimator must always satisfy this. For any other β, we have:

$$\begin{aligned}(X - A\beta)^T(X - A\beta) &= (X - A\hat{\beta} + A\hat{\beta} - A\beta)^T(X - A\hat{\beta} + A\hat{\beta} - A\beta) \\ &= (X - A\hat{\beta})^T(X - A\hat{\beta}) + (\hat{\beta} - \beta)^T A^T A(\hat{\beta} - \beta) + 2(X - A\hat{\beta})^T A(\hat{\beta} - \beta) \\ &\geq (X - A\hat{\beta})^T(X - A\hat{\beta})\end{aligned}$$

(Notice that $(\hat{\beta}-\beta)^T A^T A(\hat{\beta}-\beta)$ is always positive, and $(X-A\hat{\beta})^T A(\hat{\beta}-\beta) = 0$ by the normal equation.) Thus, any estimate that satisfies the normal equation minimizes the squared residuals.

¿From a geometric point of view, we may think of $A\beta$ as the image of the linear transformation, $A : R^k \to R^n$. Then, minimizing the squared residuals is equivalent to finding the point in the image closest to X. This is the projection of X onto $A\beta$; this projection is unique. Note that the residuals must be orthogonal to $A\beta$-space. If A is one-to-one, then there is exactly one β that maps to each point in the space; the β that maps to the projection is our estimate, $\hat{\beta}$. Otherwise, there is an entire flat of R^k that is mapped to the projection, and there is not a unique $\hat{\beta}$, even though there is a unique projection.

If A is of rank k, then we say A is of *full rank*, we know that $(A^T A)^{-1}$ exists, and the least squares estimate is:

$$\hat{\beta} = (A^T A)^{-1} A^T X$$

In this case, the fitted values (which are the estimates of $E(X)$ and are the projection onto $A\beta$-space), are:

$$\hat{X} = A\hat{\beta} = A(A^T A)^{-1} A^T X$$

$H = A(A^T A)^{-1} A^T$ is the *projection matrix* onto $Im(A)$. Some of its properties include:

- H is idempotent; that is $H^2 = H$.
- The null space of H is the space orthogonal to A.
- H is symmetric.

The estimates of the parameters and the fitted values are unbiased:

$$\begin{aligned} E(\hat{\beta}) &= E((A^T A)^{-1} A^T X) \\ &= (A^T A)^{-1} A^T E(X) \\ &= (A^T A)^{-1} A^T A\beta \\ &= \beta \\ E(A\hat{\beta}) &= AE(\hat{\beta}) \\ &= A\beta \\ &= E(X) \end{aligned}$$

Under the additional assumption that the errors are uncorrelated and have equal variance (that is, $Var(\epsilon) = \sigma^2 I$),

$$\begin{aligned} Var(\hat{\beta}) &= Var((A^T A)^{-1} A^T X) \\ &= (A^T A)^{-1} A^T Var(X) A (A^T A)^{-1} \\ &= \sigma^2 (A^T A)^{-1} \end{aligned}$$

If A is not of full rank, then we say β is *not identifiable*, since A is a many-to-one map. That is, different values of β lead to the same distribution of the X. Suppose $rank(A) = r$. Without loss of generality, assume that the first r columns of A span the column space of A. Let $B = (A_1, ..., A_r)$. Then there is some γ such that $A\beta = B\gamma$. Since B is invertible, γ is identifiable. There are multiple possible choices for B, but given B, the estimate of γ is unique.

Theorem 6.1 Gauss-Markov Theorem *Suppose $X = A\beta + \epsilon$ with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 I$. Then the least squares estimator is the minimum variance unbiased* linear *estimator.*

Proof Let $\phi = c^T \beta$ where c is a k-dimensional vector. $\hat{\beta}$ is the minimum variance linear unbiased estimator if and only if, for any other unbiased estimator, $\hat{\phi} = b^T X$, $Var(c^T \hat{\beta}) \leq Var(\hat{\phi})$. Since $\hat{\beta}$ is unbiased for β, $c^T \hat{\beta}$ is unbiased for ϕ. In addition, for any $\beta \in R^k$,

$$\begin{aligned} c^T \beta &= E(\hat{\phi}) \\ &= b^T E(X) \\ &= b^T A\beta \end{aligned}$$

This means that $c^T = b^T A$.

First, suppose A is of full rank. Then,

$$\begin{aligned} Var(c^T\hat{\beta}) &= c^T Var(\hat{\beta})c \\ &= \sigma^2 c^T (A^T A)^{-1} c \\ &= \sigma^2 b^T A (A^T A)^{-1} A^T b \\ Var(\hat{\phi}) &= b^T Var(X) b \\ &= \sigma^2 b^T b \\ Var(\hat{\phi}) - Var(c^T\hat{\beta}) &= \sigma^2 b^T (I - A(A^T A)^{-1} A^T) b \end{aligned}$$

The matrix $I - A(A^T A)^{-1} A^T$ maps any vector to its component that is orthogonal to the image of A. Thus, this matrix is symmetric and idempotent, and

$$\begin{aligned} Var(\hat{\phi}) - Var(c^T\hat{\beta}) &= \sigma^2 b^T (I - A(A^T A)^{-1} A^T) b \\ &= \|(I - A(A^T A)^{-1} A^T) b\|^2 \\ &\geq 0 \end{aligned}$$

Thus, no other linear unbiased estimator has smaller variance, and when A has full rank, $\hat{\beta}$ is the minimum variance linear unbiased estimator.

Now, suppose $rank(A) = r < k$. In this case, $\phi = c^T\beta$ is not identifiable for some c. If $c^T = b^T A$ for some b, then ϕ is estimable, using $b^T(A\hat{\beta})$. We show that this is the best linear unbiased estimator. Since $b \in R^n$, we we may write $b = a + (b - a)$, where $a \in Im(A)$ and, therefore, $b - a$ is orthogonal to $Im(A)$. Then,

$$E(b^T X) = a^T A\beta + (b-a)^T A\beta = a^T A\beta$$

Since the original estimator is unbiased, $a^T A\beta = c^T\beta$, and $a^T X$ is also a linear unbiased estimator. In addition, this estimator has a smaller (or equal) variance:

$$\begin{aligned} Var(a^T X) &= a^T Var(X) a \\ &= \sigma^2 a^T a \\ Var(b^T X) &= b^T Var(X) b \\ &= \sigma^2 (a^T + (b-a)^T)(a + (b-a)) \\ &= \sigma^2 a^T a + \sigma^2 (b-a)^T (b-a) + 0 \\ &\geq Var(a^T X) \end{aligned}$$

We show that the estimator found through projection is always the least squares estimator. We know that the least squares estimator is $c^T\hat{\beta}$, where $\hat{\beta}$ satisfies the normal equation, $A^T A\hat{\beta} = A^T X$, even though we cannot solve this equation. Since $c^T = b^T A$,

$$\begin{aligned} c^T\hat{\beta} &= b^T A\hat{\beta} \\ &= (a^T + (b-a)^T) A\hat{\beta} \\ &= a^T A\hat{\beta} \end{aligned}$$

Since $a \in Im(A)$, we may write $a^T A\hat{\beta} = a^T X$, and we see that $c^T\hat{\beta} = a^T X$. Thus, the least squares estimator is the minimum variance unbiased estimator whenever ϕ is estimable. ∎

We may rewrite:

$$\begin{aligned}\epsilon^T\epsilon &= (X - A\beta)^T(X - A\beta) \\ &= (X - A\hat{\beta} + A(\hat{\beta} - \beta))^T(X - A\hat{\beta} + A(\hat{\beta} - \beta)) \\ &= (X - A\hat{\beta})^T(X - A\hat{\beta}) + (\hat{\beta} - \beta)^T A^T A(\hat{\beta} - \beta) + 0\end{aligned}$$

Suppose $Var(\epsilon) = \sigma^2 I$. The first term is the sum of squared observed residuals, and can be calculated from the data. Consider an orthonormal change of coordinates (that is, a rotation), so that $Im(A)$ coincides with the first $r = rank(A)$ coordinates (remember, this need not be equal to the number of parameters!). Then, this maps ϵ to η, where $\eta_1, ..., \eta_r$ represent $A(\hat{\beta} - \beta)$ and $\eta_{r+1}, ..., \eta_n$ represent $X - A\hat{\beta}$. Since this is a rotation, lengths and angles are preserved, and $\eta^T\eta = \epsilon^T\epsilon$ and $Var(\eta) = \sigma^2 I$. In particular,

$$\begin{aligned}E((X - A\hat{\beta})^T(X - A\hat{\beta})) &= E((\eta_{r+1}, ..., \eta_n)^T(\eta_{r+1}, ..., \eta_n)) \\ &= Var((\eta_{r+1}, ..., \eta_n)) \\ &= (n - r)Var(\eta_{r+1}) = (n - r)\sigma^2\end{aligned}$$

Thus, $\hat{\sigma}^2 = \frac{1}{n-r}(X - A\hat{\beta})^T(X - A\hat{\beta})$ is an unbiased estimator of σ^2. Also, $n - r$ is the number of degrees of freedom in $\hat{\epsilon}$ and r is the degrees of freedom used up in estimation.

6.1 Generalized Least Squares

Suppose $X = A\beta + \epsilon$ with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2\Sigma$, with Σ known. (The simplest case is where Σ is a non-identity diagonal matrix, meaning that errors are uncorrelated, but have different variances. Autocorrelation in the errors also requires the use of generalized least squares.) Since Σ is a covariance matrix, it must be symmetric and positive definite, and we may write $\Sigma = PP^T$ for some matrix P. Let $z = P^{-1}X = P^{-1}A\beta + P^{-1}\epsilon$ and $\eta = P^{-1}\epsilon$. Then, $Pz = A\beta + P\eta$, and $z = P^{-1}A\beta + \eta$, and $Var(\eta) = P^{-1}Var(\epsilon)(P^{-1})^T = \sigma^2 I$. In fact,

$$\begin{aligned}(z - P^{-1}A\beta)^T(z - P^{-1}A\beta) &= (P^{-1}X - P^{-1}A\beta)^T(P^{-1}X - P^{-1}A\beta) \\ &= (X - A\beta)^T\Sigma^{-1}(X - A\beta)\end{aligned}$$

So we can minimize the sum of squared residuals in the original model, weighted by the inverse of the covariance matrix.

6.2 The Normal Assumption

Up to now, we have assumed nothing about the distribution of ϵ. For some further results, we assume that $\epsilon \sim Normal(0, \sigma^2 I)$. If A is also of full rank, then $\hat{\beta} \sim Normal(\beta, \sigma^2(A^T A)^{-1})$ and, for any c, $c^T\hat{\beta} \sim Normal(c^T\beta, \sigma^2 c^T(A^T A)^{-1}c)$. Using the rotation from before, we know that $(X - A\hat{\beta})^T(X - A\hat{\beta}) = \sum_{i=k+1}^{n} \eta_i^2$. Because the η_i are now independent normal with variance σ^2, we know that $\frac{1}{\sigma^2}(X - A\hat{\beta})^T(X - A\hat{\beta}) \sim \chi^2_{n-k}$.

Since a distribution is specified, we may also write down a likelihood function. Let $\theta = (\beta_1, ..., \beta_k, \sigma^2)$. Then,

$$\begin{aligned} lik(\theta, A, y) &= P_\theta(y \mid A) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(y - A\beta)^T(y - A\beta)\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(\frac{1}{2\sigma^2}(-y^T y + 2\beta^T A^T y - \beta^T A^T A\beta)\right) \\ &= C(\sigma^2, \beta) \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \sum_{j=1}^{n} \frac{\beta_j}{\sigma^2} t_j(y)\right) \end{aligned}$$

where $C(\sigma^2, \beta)$ depends only on θ and $t_j(y)$ is the j^{th} component of $A^T y$. Thus, this model is in the exponential family with complete sufficient statistics $\sum y_i^2$ and $A^T y$, corresponding to $\beta_1/\sigma^2, ..., \beta_k/\sigma^2, 1/2\sigma^2$. Since this parameter space contains a $k+1$-dimensional ball, any unbiased estimator that is a function of these statistics is the minimum variance unbiased estimator. $\hat{\beta}$ is clearly a function of these sufficient statistics, and $\hat{\sigma}^2$ is a function of $\hat{\beta}$ and $\sum y_i^2$. (It is also the maximum likelihood estimator.)

6.3 Estimation with Restrictions on β

Suppose $y = A\beta + \epsilon$ with the linear restriction $H^T\beta = 0$, with $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2 I$, and H an $k \times q$-matrix. Notice that we may assume that H is of full rank; otherwise, some of the restrictions are redundant, and we may delete them with no loss of information. Also, if the restrictions are of the form $H\beta = c$ for some constant, then we may reparameterize β to absorb c. However, this method does not apply to any nonlinear restrictions, where β is restricted to a subset of R^k that is not a subspace.

In this case, the space of possible β-values has dimension $k - q$. Now, we must minimize the sum of squares subject to the constraint; using the Lagrange multiplier, this gives us the following system of linear equations:

$$\begin{aligned} A^T A\beta - A^T y + H\lambda &= 0 \\ H^T\beta &= 0 \\ \begin{pmatrix} A^T A & H \\ H^T & 0 \end{pmatrix} \begin{pmatrix} \beta \\ \lambda \end{pmatrix} &= \begin{pmatrix} A^T y \\ 0 \end{pmatrix} \end{aligned}$$

If A is of full rank, then $\Sigma = \begin{pmatrix} A^TA & H \\ H^T & 0 \end{pmatrix}$ has an inverse of the form $\Sigma^{-1} = \begin{pmatrix} P & Q^T \\ Q & R \end{pmatrix}$, with

$$\begin{aligned} P &= (A^TA)^{-1} - (A^TA)^{-1}H(H^T(A^TA)^{-1}H)^{-1}H^T(A^TA)^{-1} \\ Q^T &= (H^T(A^TA)^{-1}H)^{-1}H^T(A^TA)^{-1} \\ R &= -(H^T(A^TA)^{-1}H)^{-1} \end{aligned}$$

Then, $\hat{\beta} = PA^Ty$ and $\hat{\lambda} = QA^Ty$. By definition of P, Q, and R:

$$\begin{aligned} PA^TA + QH &= I \\ PH^T &= 0 \\ QA^TA + RH^T &= 0 \\ Q^TH &= I \end{aligned}$$

We may use these identities (along with $H\beta = 0$) to show that $\hat{\beta}$ is still unbiased and $E(\hat{\lambda}) = E(QA^Ty) = QA^TA\beta = 0$. Furthermore, $Var(\hat{\beta}) = \sigma^2 P$, $Var(\hat{\lambda}) = -\sigma^2 R$, and the two are uncorrelated.

If A does not have full rank, then some (or all) of the restrictions may be used to identify β while the rest actually restrict it. Suppose $H = (H_1, H_2)^T$, where H_1 are conditions that exactly identify β (along with A). Then $A^TA + H_1^TH_1$ is invertible. If we replace A^TA above by $A^TA + H_1^TH_1$, then everything goes through as before.

7 Maximum Likelihood

Let $\theta \in \Theta \subset R^k$ be a parameter of interest. Suppose $X = (x_1, ..., x_n)$ are independent and identically distributed from $f(x; \theta_0)$, where θ_0 is the true parameter value. Then, we have the likelihood function for θ:

$$\begin{aligned} lik(\theta \mid X) &= f(X \mid \theta) \\ &= \prod_{i=1}^{n} f(x_i \mid \theta) \end{aligned}$$

This is a function of θ. In *maximum likelihood estimation*, we choose $\hat{\theta}$ to maximize the likelihood function for the observed data. To do this, we generally maximize the *log likelihood*, $l(\theta)$, because the natural logarithm is a monotonically increasing function (and the log likelihood is usually easier to work with). To maximize this, we check both points where the derivative is zero and the boundaries of Θ.

Definition $\hat{\theta}_n \to \theta_0$ *in probability* if for all $\epsilon > 0$, $P_{\theta_0}(|\hat{\theta}_n - \theta_0| > \epsilon) \to 0$ as $n \to \infty$. This is also called *weak convergence*.

Definition $\hat{\theta}_n \to \theta_0$ *almost surely* if $P_{\theta_0}(\omega \mid \lim \hat{\theta}_n(\omega) = \theta_0) \to 1$ as $n \to \infty$. This is also called *strong convergence*.

Definition If $\hat{\theta}_n(X) \to \theta_0$ in probability, then $\hat{\theta}_n$ is *weakly consistent*. If $\hat{\theta}_n \to \theta_0$ almost surely, then $\hat{\theta}_n$ is *strongly consistent*. In both cases, $\hat{\theta}_n = \theta_0 + o_p(1)$; that is, the difference between the estimate and the true value is of order smaller than 1.

Definition If $P_{\theta_0}(|\hat{\theta}_n - \theta_0| < M) \to 1$ for a fixed number M, then $\hat{\theta}_n$ is *bounded in probability*, and we write $\hat{\theta}_n = \theta_0 + O_p(1)$, since their difference is of order 1.

Theorem 7.1 *The maximum likelihood estimator is weakly consistent if the sample is independent and identically distributed and if the number of parameters is constant as the sample size grows.*

Proof *(Sketch.)* Define $Z(\theta) = E_{\theta_0}(l(x_1, \theta))$, where θ_0 is the true parameter value. Define $\hat{Z}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(x_i, \theta)$. By the law of large numbers, $\hat{Z}_n(\theta) \to Z(\theta)$ for all θ.

We first show that Z is maximized at θ_0.

$$\begin{aligned} Z(\theta_0) - Z(\theta) &= \int \log(f(x,\theta_0)) f(x,\theta_0)\, dx - \int \log(f(x,\theta)) f(x,\theta_0)\, dx \\ &= \int \log(\frac{f(x,\theta_0)}{f(x,\theta)}) f(x,\theta_0)\, dx \\ &= \int -\log(\frac{f(x,\theta)}{f(x,\theta_0)}) f(x,\theta_0)\, dx \end{aligned}$$

Since the negative logarithm is convex, we may apply Jensen's Inequality, which states that if f is a convex function then $E(f(X)) \geq f(E(X))$, with strict inequality if the function is not a line.

$$\begin{aligned} Z(\theta_0) - Z(\theta) &= \int -\log(\frac{f(x,\theta)}{f(x,\theta_0)}) f(x,\theta_0)\, dx \\ &\geq -\log(\int \frac{f(x,\theta)}{f(x,\theta_0)} f(x,\theta_0)\, dx) \\ &= -\log \int f(x,\theta)\, dx \\ &= -\log(1) = 0 \end{aligned}$$

In this case, strict inequality holds if $f(x,\theta) \neq f(x,\theta_0)$ on a set of positive measure. Thus, $Z(\theta)$ is maximized at θ_0.

We consider the case where the parameter space is finite; that is, $\Theta = \{\theta_0, \theta_1, ..., \theta_k\}$. Since there are finitely many parameter values, the convergence of $\hat{Z}_n(\theta_j) \to Z(\theta_j)$ must be uniform. Let $\epsilon = min\{Z(\theta_1) - Z(\theta_0), ..., Z(\theta_k) - Z(\theta_0)\}$. Without loss of generality, assume that $Z(\theta_1) - Z(\theta_0) = \epsilon$. Then we

may choose N so that the probability of the $\hat{Z}_n(\theta_j)$ being far enough away from their true values to make the minimum $\hat{Z}_n(\theta_j)$ different from zero is as small as we want. Thus, the probability that $\hat{\theta}_n$ equals θ_0 goes to 1 as $n \to \infty$.

For more general cases, some useful assumptions include:

- $\lim_{m\to\infty} E(sup_{|\theta-\theta_0|>m} \frac{1}{n}\sum l(x_i,\theta) - \frac{1}{n}\sum l(x_i,\theta_0)) < 0$. That is, far away from θ_0, the likelihood is smaller.
- The likelihood function is continuous with respect to θ.
- Singularities in $lik(\theta)$ do not depend on θ.
- The parameter space is finite-dimensional and compact. ∎

Theorem 7.2 *Under additional regularity conditions, the maximum likelihood estimator is asymptotically normal, so that* $\sqrt{n}(\hat{\theta} - \theta_0) \sim Normal(0, i_{\theta_0}^{-1})$, *and* $\hat{\theta} \sim Normal(\theta_0, I_n(\theta_0)^{-1})$.

Proof *(Sketch.)* Let $D(\theta) = \frac{\partial}{\partial\theta} \log f(x,\theta)|_{\theta=\theta_0}$ and $D^2(\theta) = \frac{\partial^2}{\partial\theta^2} \log f(x,\theta)|_{\theta=\theta_0}$. Using a Taylor expansion, we find that:

$$0 = D(\hat{\theta}) \approx D(\theta_0) + D^2(\theta_0)(\hat{\theta} - \theta_0)$$

Then, $\sqrt{n}(\hat{\theta}-\theta_0) \approx \sqrt{n}\frac{1}{D^2(\theta_0)}D(\theta_0)$. By the Law of Large Numbers, $-\frac{1}{n}D^2(\theta_0) \to i(\theta_0)$. Since $\frac{1}{\sqrt{n}}D(\theta_0)$ is a sum of independent and identically distributed random variables, each with mean 0 and variance $i(\theta)$, the sum converges to the normal distribution, $\frac{1}{\sqrt{n}}D(\theta_0) \sim Normal(0, i_\theta)$. Thus, $\sqrt{n}(\hat{\theta} - \theta_0) \sim Normal(0, I_{\theta_0}^{-1})$ asymptotically. ∎

Definition An estimator, $\hat{\theta}_n$, is *self-consistent*, or *invariant*, if $g(\hat{\theta}_n)$ is consistent for $g(\theta)$ for any function g.

Note that the maximum likelihood estimator is self-consistent, because the θ that maximizes the likelihood must be mapped to a value $g(\theta)$ that maximizes the reparameterized likelihood. In the regular case, $\sqrt{n}(g(\hat{\theta}_{mle}) - g(\theta_0)) \to Normal(0, (g'(\theta_0))^2 i(\theta_0)^{-1})$ in distribution. (This also holds in the multivariate case, though the variance is now written as $g'(\theta_0)I(\theta_0)^{-1}g'(\theta_0)$, which is also the Cramer-Rao lower bound for estimating $g(\theta)$.) In finite samples, the Fisher information is (as always) approximated by the information at $\hat{\theta}$.

This can be applied to variance-stabilizing transformations. Suppose that $x \sim [\mu, \sigma^2]$. Then, $Var(g(\bar{x})) \approx (g'(\mu))^2\sigma^2/n$. If we may find a g such that $(g'(\mu))^2\sigma^2$ is a constant, this may be useful in estimation, particularly if you are near a parameter value with unstable variance.

7.1 Maximum Likelihood and the Exponential Family

Suppose $f(x,\theta) = \exp(a(\theta)b(x) + c(\theta) + d(x)), \theta \in R$, and that $x_1, ..., x_n$ are a random sample from this distribution. Then,

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log f(x_i,\theta) &= b(x_i)a'(\theta) + c'(\theta) \\
0 &= E(\frac{\partial}{\partial\theta}\log f(x_i,\theta)) = a'(\theta)E(b(x_i)) + c'(\theta) \\
E_{\theta_0}(b(x_i)) &= -c'(\theta_0)/a'(\theta_0) = \mu(\theta_0) \\
Var(\frac{\partial}{\partial\theta}\log f(x_i,\theta)) &= (a'(\theta))^2 Var(b(x_i))
\end{aligned}$$

Notice that the last equation equals the Fisher information. We may use the alternative definition of the Fisher information to solve for $Var(b(x_i))$:

$$\begin{aligned}
i(\theta) &= -E(\frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta)) \\
&= -E(b(x_i))a''(\theta) - c''(\theta) \\
&= -\mu(\theta)a''(\theta) - c'(\theta) \\
-\mu(\theta)a''(\theta) - c'(\theta) &= (a'(\theta))^2 Var(b(x_i)) \\
Var(b(x_i)) &= \frac{-a''(\theta)\mu(\theta) - c''(\theta)}{(a'(\theta))^2}
\end{aligned}$$

In this case, the maximum likelihood estimator is the mean of $b(x_i)$. Notice that the second derivative of the log likelihood function is always negative if $a'(\theta) \neq 0$, because the variance of $b(x_i)$ is positive. This gives a global maximum for the exponential family. (This method also extends to vector-valued θ.)

7.2 Restricted Maximum Likelihood Estimation

Suppose we want to estimate θ subject to a restriction $h(\theta) = (h_1(\theta), ..., h_r(\theta)) = 0$ (where the dimension of the restriction is less than the dimension of the parameter space, and the restriction may be non-linear). Let $H_\theta = (\frac{\partial h_i}{\partial\theta_j})|_\theta$. Then, Lagrange multipliers give us the following system of equations:

$$\begin{aligned}
D(l(\theta,x)) - H_\theta^T\lambda &= 0 \\
h(\theta) &= 0
\end{aligned}$$

Let $\hat{\theta}$ be the unrestricted estimate and $\tilde{\theta}$ be the restricted estimate. We assume that $\hat{\theta}, \tilde{\theta}$, and θ_0 are all sufficiently close that we may use the following Taylor expansions:

$$\begin{aligned}
h(\tilde{\theta}) &= h(\theta_0) + H_\theta(\tilde{\theta} - \theta_0) \\
Dl(\tilde{\theta}) &= Dl(\theta_0) + (\tilde{\theta} - \theta_0)D^2 l(\theta_0, x) \\
Dl(\tilde{\theta}) &\approx Dl(\hat{\theta}, x) = 0 \\
\frac{1}{n}D^2 l(\theta_0, x) &\approx i(\theta_0)
\end{aligned}$$

Multiplying everything by $\sqrt{n}$ gives us a system of equations, which can also be written in matrix form:

$$\begin{aligned}
\sqrt{n}(\tilde{\theta}-\theta_0)i(\theta_0)+\frac{1}{\sqrt{n}}H_\theta^T\tilde{\lambda} &= \frac{1}{\sqrt{n}}Dl(\theta_0,x)\\
H_\theta(\sqrt{n})(\tilde{\theta}-\theta_0) &= 0\\
\begin{pmatrix} i(\theta_0) & H^T \\ H & 0 \end{pmatrix}\begin{pmatrix} \sqrt{n}(\tilde{\theta}-\theta_0) \\ \frac{1}{\sqrt{n}}\tilde{\lambda} \end{pmatrix} &= \begin{pmatrix} \frac{1}{\sqrt{n}}Dl(\theta_0,x) \\ 0 \end{pmatrix}
\end{aligned}$$

We know that this square matrix has an inverse of the form $\begin{pmatrix} P & Q^T \\ Q & R \end{pmatrix}$. Thus, because $\frac{1}{\sqrt{n}}P\,Dl(\theta_0,x)$ is asymptotically normal, $\sqrt{n}(\tilde{\theta}-\theta_0)$ is asymptotically normal as well.

7.3 Newton's Method

In practice, closed forms solutions may not exist. Instead, we use *Newton's Method* to estimate the point at which the maximum occurs. Suppose we know $l(\theta)$. Let $D_\theta = (\frac{\partial}{\partial\theta_1},...,\frac{\partial}{\partial\theta_n})$. We wish to solve the equation $D_\theta l = 0$, since either a solution of this equation (or a boundary) is the maximum likelihood estimator. For this, suppose the solution is the correct estimator. Suppose we have an estimate, $\theta^{(n)}$, close enough to the true maximum, $\hat{\theta}$ to admit a Taylor expansion. Then,

$$\begin{aligned}
0 &= D_\theta l(\hat{\theta})\\
&\approx D_\theta l(\theta^{(n)}) + D_\theta^2 l(\theta^{(n)})(\hat{\theta}-\theta^{(n)})
\end{aligned}$$

If $D_\theta^2 l(\theta^{(n)})$ is invertible, then we may solve:

$$\theta^{(n+1)} \approx \theta^{(n)} - D_\theta^2 l(\theta^{(n)})^{-1} D_\theta l(\theta^{(n)})$$

This process continues until the estimates converge.

If the initial guess $\theta^{(0)}$ is relatively good, then the matrix $D_\theta^2 l(\theta^{(0)})^{-1}$ can be used in every step, instead of being recalculated (since inverting matrices is a very slow process). In addition, since $-D_\theta^2 l(\theta) \approx I(\theta)$ (the latter is the expectation of the former), we may also use the inverse Fisher's information of the initial estimate, if that is easier to calculate. In this case, the updating formula is:

$$\theta^{(n+1)} = \theta^{(n)} + I(\theta^{(0)})^{-1}(\hat{\theta}-\theta^{(n)})$$

None of these variations of Newton's method are guaranteed to converge to the correct value (they may find a local but not global maximum, for example).

8 Estimating equations

Definition The equation $g(\theta,x)=0$ is an *estimating equation* if $E(g(\theta,x))=0$ for all possible θ.

Given data, we may use an estimating equation to estimate a parameter by plugging in the observed data and solving for θ. Note that the derivative of the log likelihood is an estimating equation (under sufficiently regular conditions).

As an example, we consider a one-dimensional maximum likelihood estimator for a multinomial distribution with k cells and an asymptotically equivalent one defined by an estimating equation. First, for the maximum likelihood estimator, we have

$$\begin{aligned} g(\theta, s_1, ..., s_k) &= \frac{\partial}{\partial \theta} l(\theta) \\ &= n \sum_{i=1}^{k} s_i \frac{\pi_i'(\theta)}{\pi_i(\theta)} \end{aligned}$$

(This is an estimation equation, since substituting $\pi_i(\theta)$ for s_i gives zero.) Using a Taylor expansion about the true parameter value, we have:

$$g(\hat{\theta}, s_1, ..., s_k) = g(\theta_0, \pi_1(\theta), ..., \pi_k(\theta)) + (\hat{\theta}) \frac{\partial g}{\partial x_0}|_{x_0=\theta_0} + \sum_{i=1}^{k} (s_i - \pi_i(\theta_0)) \frac{\partial g}{\partial x_i}|_{x_i=\pi_i(\theta_0)} + o_p(1)$$

$$\begin{aligned} \frac{\partial g}{\partial x_0} &= n \sum_{i=1}^{k} s_i (\frac{\pi_i''(\theta)}{\pi_i(\theta)} + (\frac{\pi_i'(\theta)}{\pi_i(\theta)})^2) \rightarrow -ni(\theta) \\ \frac{\partial g}{\partial x_i} &= n \frac{\pi_i'(\theta)}{\pi_i(\theta)} \\ 0 &= 0 + (\hat{\theta} - \theta_0)(-ni(\theta_0)) + \sum_{i=1}^{k} (s_i - \pi_i) n \frac{\pi_i'(\theta_0)}{\pi_i(\theta_0)} \\ \hat{\theta} - \theta_0 &= \sum_{i=1}^{k} (s_i - \pi_i) \frac{\pi_i'(\theta)}{i(\theta)\pi_i(\theta)} \end{aligned}$$

Any estimator that asymptotically has the same form (in particular, that is a multiple of $\frac{\pi_i'(\theta)}{i(\theta)\pi_i(\theta)}$) will asymptotically equal the maximum likelihood estimator, and therefore will also be efficient. (The properties in finite samples will differ, though.) In particular, the minimal chi-square estimator, in which $G(\theta) = \sum_{i=1}^{k} \frac{1}{n_i} (n_i - n\pi_i(\theta))^2$ is minimized gives the estimation equation $g(\theta, s_1, ..., s_k) = \frac{\partial}{\partial \theta} G(\theta) = 2n \sum_{i=1}^{k} \frac{\pi_i(\theta)\pi_i'(\theta)}{s_i}$ which has an asymptotic ratio of derivatives $-\frac{\partial g}{\partial x_i} / \frac{\partial g}{\partial x_0} = \frac{\pi_i'(\theta)}{\pi_i(\theta) i(\theta)}$, and is therefore asymptotically efficient.

9 The Jackknife Method

Given data, $X_1, ..., X_n$ from an unknown distribution, F, suppose we have an estimator, $\hat{\theta}_n = \hat{\theta}_n(x_1, ..., x_n)$, of a real-valued $\theta_0(F)$. We assume that the x_i are exchangeable, and that we may write:

$$E(\hat{\theta}_n) = \theta_0 + \frac{a_1(\theta_0)}{n} + \frac{a_2(\theta_0)}{n^2} + ...$$

We define $\hat{\theta}_{n-1,i}$ as the estimator based on the $n-1$ points $X_1, ..., X_{i-1}, X_{i+1}, ..., X_n$, and $\hat{\theta}_{n-1,.}$ as the average of $\hat{\theta}_{n-1,1}, ..., \hat{\theta}_{n-1,n}$. The *jackknife estimate of bias* or *Quenouille's estimate of bias* is:

$$\hat{bias} = (n-1)(\hat{\theta}_{n-1,.} - \hat{\theta}_n)$$

Using this, we define the *jackknife estimator* as

$$\tilde{\theta}_n = n\hat{\theta}_n - (n-1)\hat{\theta}_{n-1,.}$$

Then,

$$\begin{aligned} E(\hat{\theta}_{n-1,.}) &= \theta_0 \frac{a_1(\theta_0)}{n-1} + \frac{a_2(\theta_0)}{(n-1)^2} + ... \\ E(\tilde{\theta}_n) &= E(n\hat{\theta}_n - (n-1)\hat{\theta}_{n-1,.}) \\ &= n(\theta_0 + \frac{a_1(\theta_0)}{n} + \frac{a_2(\theta_0)}{(n)^2} + ...) - (n-1)(\theta_0 \frac{a_1(\theta_0)}{n-1} + \frac{a_2(\theta_0)}{(n-1)^2} + ...) \\ &= \theta_0 + \frac{-a_2(\theta_0)}{n(n-1)} + ... \end{aligned}$$

This shows that the bias of the new estimator is of a smaller order than the bias of the original estimator. In fact, the resulting bias will be zero if the only bias was in the $\frac{a_1(\theta_0)}{n}$ term. More corrections can be done, but each correction tends to increase the variance of the estimator, so there may be a tradeoff.

This method of deleting observations one at a time can also be used to estimate variance. In particular, we have the Tukey estimate of variance:

$$\hat{Var}(\hat{\theta}_n) = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{n-1,i} - \hat{\theta}_{n-1,.})^2$$

(We don't divide by n since there is so much overlapping information being used.) This estimate captures the variance of the *influence function.*

Theorem 9.1 *$E(\hat{Var}_n) \geq Var(\hat{\theta}_n)$, and this variance estimate is conservative.*

Proof *(Sketch.)* We may use an ANOVA decomposition on $\hat{\theta}_{n-1}$:

$$\begin{aligned} \hat{\theta}_n &= E(\hat{\theta}) + (\sum_{i=1}^{n} E(\hat{\theta}_n|X_i) - \mu) + (\sum_{i<j} E(\hat{\theta}_n|X_i, X_j) - \sum_i E(\hat{\theta}_n|X_i) + \mu) + ... + ((\hat{\theta}_n + ... + (-1)^n E(\hat{\theta}) \\ &= \mu + \frac{1}{n}\sum_i \alpha_i + \sum_{i<j} \frac{1}{n^2}\frac{1}{n^n}\beta_{ij} + ... + \eta_{12...n} \end{aligned}$$

Notice that this is a decomposition into 2^n random variables, $\mu, \alpha_1, ..., \alpha_n, \beta_1 1, ...\eta_{12...n}$. It turns out that:

- Each of the random variables is a function only of the X_i referred to in the subscripts (that is, α_i depends only on X_i, β_{ij} depends on X_i and X_j and nothing else, and so on).

- Each of the random variables has expectation zero if it is conditioned on zero up to all but one of the defining X_i's.
- The random variables are pairwise uncorrelated (and therefore have expected pairwise products equal to 0).

This allows us to calculate the variance as the sum of the variances of the 2^n random variables. Notice that each of the α's has the same variance (and so on). This means we may count up the variables:

$$Var(\hat{\theta}_{n-1}) = \frac{1}{n-1}Var(\alpha_i)+\binom{n-2}{1}\frac{1}{2(n-1)^3}Var(\beta_{ij})+\binom{n-2}{2}\frac{1}{3(n-1)^5}Var(\gamma_{ijk})+...$$

This is the true variance of the jackknife estimator. This allows us to calculate the expectation of the estimated variance as well:

$$\begin{aligned} E(\sum_{i=1}^{n}(\hat{\theta}_{n-1,i} - \hat{\theta}_{n-1,\cdot})^2) &= \frac{1}{n}\sum_{i<j}E((\hat{\theta}_{n-1,i} - \hat{\theta}_{n-1,j})^2) \\ &= \frac{Var(\alpha)}{n-1} + \binom{n-2}{1}\frac{Var(\beta)}{(n-1)^3} + \binom{n-2}{2}\frac{Var(\gamma)}{(n-1)^5} + ... \end{aligned}$$

Note that the true variance is less than the estimated variance. ∎

10 Confidence Regions

Definition A *pivotal quantity* is a random quantity whose distribution does not depend on any unknown parameters. The pivotal quantity itself may depend on unknown parameters.

Definition A $1-\alpha$ *confidence region*, $C_\alpha(x) \subset \Theta$ is a region that satisfies:

$$\inf_\theta P_\theta(\theta \in C_\alpha(x)) \geq 1-\alpha$$

Note that $C(x)$ is a random region that tries to capture the true (fixed) θ. $1-\alpha$ is called the *confidence coefficient*.

We hope that the confidence regions are as small as possible.

If $C(x)$ is based on a pivotal quantity, then $P_\theta(\theta \in C_\alpha(x))$ does not depend on any unknown parameters, and the infimum is no longer necessary. Note that quantities might be pivotal only asymptotically; for example, $\sqrt{\frac{n}{i(\theta)}}(\hat{\theta} - \theta_0)$ is pivotal asymptotically, because its distribution is asymptotically standard normal.

11 Hypothesis Testing

Suppose we have a parameter $\theta \in \Theta$ where we have a partition $\Theta = \Omega \cup \Omega^C$. We also observe a point, x, in the sample space, S. Our *null hypothesis* (H_0) is that $\theta \in \Omega$, and our *alternative hypothesis* (H_A) is that $\theta \in \Omega^C$. We choose a *critical set*, $R \subset S$, and accept H_0 if and only if $x \in R^C$ (equivalently, reject H_0 if and only if $x \in R$). We define *Type I Error* as $P_\theta(R)$ when $\theta \in \Omega$, and *Type II Error* as $P_\theta(R^C)$ when $\theta \in \Omega$. The *size* of the test is $\max P_\theta(R)$ when $\theta \in \Omega$ (this is the maximum Type I Error). The *power* of the test is $1 - P_\theta(R^c)$ when $\theta \in \Omega^C$ (this is one minus the Type II Error). Ideally, we want both types of error to be small. In general, we choose a size, α, and then try to find the most powerful test (critical region) with that size.

Definition A test of size α with a power function that is uniformly no larger than that of any other test of size α (or less) is the *uniformly most powerful* (UMP) test.

Theorem 11.1 Neyman-Pearson Lemma. *Suppose we have the simple hypotheses, $H_0 : \theta = \theta_0$ and $H_A : \theta = \theta_1$. Suppose there is a critical region, R^*, such that $R^* = \{X \in S : f_{\theta_1}(x) \geq k f_{\theta_0}(x), k > 0\}$ and $P_{\theta_0}(R^*) = \alpha$. Then, for any other R with $P_{\theta_0}(R) \leq \alpha$, $P_{\theta_1}(R^*) \geq P_{\theta_1}(R)$. That is, a test of simple hypotheses based on the ratio of the likelihoods is the most powerful.*

Proof

$$\begin{aligned}
P_{\theta_1}(R^*) - P_{\theta_1}(R) &= \int_{R^* \cap R} f_{\theta_1}(x)dx + \int_{R^* \cap R^C} f_{\theta_1}(x)dx - (\int_{R^* \cap R} f_{\theta_1}(x)dx + \int_{R^{*C} \cap R} f_{\theta_1}(x)dx) \\
&= \int_{R^* \cap R^C} f_{\theta_1}(x)dx - \int_{R^{*C} \cap R} f_{\theta_1}(x)dx \\
&\geq \int_{R^* \cap R^C} k f_{\theta_0}(x)dx - \int_{R^{*C} \cap R} k f_{\theta_0}(x)dx) \\
&= \int_{R^*} k f_{\theta_0}(x)dx - \int_{R} k f_{\theta_0}(x)dx) \\
&= k(P_{\theta_0}(R^*) - P_{\theta_0}(R)) \\
&\geq 0
\end{aligned}$$

If the probabilities are discrete, such an R^* may not exist. In this case, we may wish to choose a more conservative region or the region with the probability closest to α. Alternately, we may create a "randomized test" which randomly assigns which hypothesis is accepted or rejected in certain cases and can achieve a size of exactly α.

Lemma 11.2 Generalized Neyman-Pearson Lemma. *Let $f, g_1, ...g_m$ be regular functions. Let $0 \leq \gamma(x) \leq 1$ and $0 \leq k_i$ for $i = 1, ..., m$. Let*

$$\phi(x) = \begin{cases} 1 & \text{if } f(x) > k_1 g_1(x) + ... + k_m g_m(x) \\ \gamma(x) & \text{if } f(x) = k_1 g_1(x) + ... + k_m g_m(x) \\ 0 & \text{if } f(x) < k_1 g_1(x) + ... + k_m g_m(x) \end{cases}$$

Then, for any ϕ_2 such that $\int \phi_2(x)g_i(x)dx \leq \int \phi(x)g_i(x)dx$ for $i = 1, ..., m$ and $0 \leq \phi_2(x) \leq 1$, $\int \phi(x)f(x)dx \geq \int \phi_2(x)f(x)dx$.

Proof Consider $\int_R (\phi(x) - \phi_2(x))(f(x) - \sum_{i=1}^m k_i g_i(x))dx$. Notice that we may write $R = A \cup B \cup C$, where $A = \{x : \phi(x) = 1\}$, $B = \{x : \phi(x) = \gamma(x)\}$, and $C = \{x : \phi(x) = 0\}$. On A, both terms in the product are positive. On B, the second term is zero. On C, both terms are negative. Thus, the integral must be positive, and we may rewrite it as:

$$\begin{aligned} 0 &\leq \int_R (\phi(x) - \phi_2(x))(f(x) - \sum_{i=1}^m k_i g_i(x))dx \\ &= \int_R (\phi(x) - \phi_2(x))f(x)dx - \int (\phi(x) - \phi_2(x))(\sum_{i=1}^m k_i g_i(x))dx \\ &= \int_R (\phi(x) - \phi_2(x))f(x)dx - \sum_{i=1}^m k_i \int_R (\phi(x) - \phi_2(x))g_i(x)dx \end{aligned}$$

Since the second term is non-negative, we must have $\int_R \phi(x)f(x)dx \geq \int_R \phi_2(x)f(x)dx$. ∎

In the case of simple hypotheses, it may happen that the likelihood ratio, $f_{\theta_1}(x)/f_{\theta_0}(x)$ is monotone in a test statistic. In this case, we may extend this most powerful test to being the most powerful test for an interval (such as one-sided tests of the mean of a normal distribution). This need not happen in general.

Definition A test such that of the (possibly composite) hypotheses, $H_O : \theta \in \Omega$ and $H_A : \theta \in \Theta - \Omega$, is a *unbiased* α-level test if $\alpha(\theta) \leq \alpha$ for all $\theta \in \Omega$ and $1 - \beta(\theta) \geq \alpha$ for all $\theta \in \Theta - \Omega$.

Definition Suppose we wish to test (possibly composite) hypotheses, $H_O : \theta \in \Omega$ and $H_A : \theta \in \Theta - \Omega$. Define the (generalized) *likelihood ratio* as:

$$\lambda(x) = \frac{\sup_{\theta \in \Theta - \Omega} f_\theta(x)}{\sup_{\theta \in \Omega} f_\theta(x)}$$

The *likelihood ratio test* rejects the null hypothesis for large values of $\lambda(x)$.

In simple cases, we may find a test statistic which is monotonically related to $\lambda(x)$, and then find a critical region for this test statistic. In more complex cases, we must use asymptotics.

In our calculation of the likelihood ratio, we use the unrestricted maximum likelihood estimator, $\hat{\theta}_n$, and the restricted maximum likelihood estimator, $\tilde{\theta}_n$. If the estimators obey regularity conditions and the sample is large enough, then $\sqrt{n}(\hat{\theta}_n - \theta_0) \sim Normal(0, B_{\theta_0}^{-1})$, where B_{θ_0} is the Fisher information. Under the null hypothesis that the restrictions hold, $\sqrt{n}(\tilde{\theta}_n - \theta_0) \sim Normal(0, P_{\theta_0})$,

where P_{θ_0} comes from the restricted maximum likelihood estimation. Then, the likelihood ratio can be written as:

$$\lambda(x) = \frac{f_{\hat{\theta}_n}(x)}{f_{\tilde{\theta}_n}(x)}$$

Theorem 11.3 *Suppose $\Omega \subset \Theta$ is a subspace. Let $d = dim(\Theta) - dim(\Omega)$ (this is the number of restrictions implicit in the null hypothesis). Then, $-2\log\lambda(x) \sim \chi^2_d$ asymptotically under the null hypothesis.*

Proof We use a Taylor expansion about $\hat{\theta}_n$:

$$\begin{aligned}
2\log\lambda(x) &= 2\log\left(\frac{f_{\hat{\theta}_n}(x)}{f_{\tilde{\theta}_n}(x)}\right) \\
&\approx 2(\log(1) - D\log(f_{\hat{\theta}_n}(x))(\tilde{\theta}_n - \hat{\theta}_n) + \frac{1}{2}(\hat{\theta}_n - \tilde{\theta}_n)^T D^2\log(f_{\hat{\theta}_n}(x))(\hat{\theta}_n - \tilde{\theta}_n)) \\
&= (\hat{\theta}_n - \tilde{\theta}_n)^T D^2\log(f_{\hat{\theta}_n}(x))(\hat{\theta}_n - \tilde{\theta}_n) \\
&\approx (\hat{\theta}_n - \tilde{\theta}_n)^T(-nB_{\theta_0})(\hat{\theta}_n - \tilde{\theta}_n)
\end{aligned}$$

Let $Y \sim Normal(0, B_{\theta_0})$. Then, under the null hypothesis, we may write $\sqrt{n}(\hat{\theta}_n - \theta_0) = B^{-1}_{\theta_0}Y$ and $\sqrt{n}(\tilde{\theta}_n - \theta_0) = P_{\theta_0}Y$, since $Var(P_{\theta_0}) = P_{\theta_0}B_{\theta_0}P_{\theta_0} = P_{\theta_0}$. Then,

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) &= \sqrt{n}(\hat{\theta}_n - \theta_0) - \sqrt{n}(\tilde{\theta}_n - \theta_0) \\
&= B^{-1}_{\theta_0}Y - P_{\theta_0}Y \\
&= (B^{-1}_{\theta_0} - P_{\theta_0})Y
\end{aligned}$$

Substituting this into the Taylor expansion, we find:

$$\begin{aligned}
2\log\lambda(x) &\approx (\hat{\theta}_n - \tilde{\theta}_n)^T(-nB_{\theta_0})(\hat{\theta}_n - \tilde{\theta}_n) \\
&= Y^T(B^{-1}_{\theta_0} - P_{\theta_0})B_{\theta_0}(B^{-1}_{\theta_0} - P_{\theta_0})Y \\
&= Y^T(B^{-1}_{\theta_0} - P_{\theta_0})Y
\end{aligned}$$

Because B_{θ_0} is symmetric and positive definite, we may write $B_{\theta_0} = A^T_{\theta_0}A_{\theta_0}$, and then $Y = A_{\theta_0}Z$, where $Z \sim Normal(0, I)$. Then, we have:

$$\begin{aligned}
2\log\lambda(x) &\approx Y^T(B^{-1}_{\theta_0} - P_{\theta_0})Y \\
&= Z^T A^T_{\theta_0}(B^{-1}_{\theta_0} - P_{\theta_0})A_{\theta_0}Z \\
&= Z^T(I - A^T_{\theta_0}P_{\theta_0}A_{\theta_0})Z
\end{aligned}$$

Note that the inner matrix is a projection matrix (and therefore idempotent). Thus, this product is distributed as the sum of r squares of independent normals, where r is the rank of the matrix (and equals the number of restrictions). Thus, $2\log\lambda(x) \sim \chi^2_r$ asymptotically. ∎

This is a generalization of an optimal test, so it may not be optimal. Also, the result above only holds in sufficiently regular cases; things go wrong if the maxima lie on the boundary.

Theorem 11.4 Wald Test. *Suppose we have restrictions on* Θ, $0 = h(\theta) = (h_1(\theta), ..., h_r(\theta))$ *and hypotheses,* $H_0 : h(\theta_0) = 0$ *and* $H_A : h(\theta_0) \neq 0$, *so that* $\Omega = \{\theta \in \Theta : h(\theta) = 0\}$. *Suppose* $dim(\Theta) = k$, *so that* $dim(\Omega) = k - r$. *Let* H_θ *be the derivative matrix of* h *and* B_{θ_0} *be the information matrix of the unrestricted estimator. Then, under regularity conditions, for large samples, under the null hypothesis,*

$$nh(\hat{\theta}_n)(H_{\theta_0}^T B_{\theta_0}^{-1} H_{\theta_0})^{-1} h(\hat{\theta}_n)^T \sim \chi_r^2$$

For practical use (since θ_0 *is unknown), we have:*

$$nh(\hat{\theta}_n)(H_{\hat{\theta}_n}^T B_{\hat{\theta}_n}^{-1} H_{\hat{\theta}_n})^{-1} h(\hat{\theta}_n)^T \sim \chi_r^2$$

Proof Using the Taylor expansion about θ_0, under the null hypothesis,

$$\begin{aligned} h(\hat{\theta}_n) &\approx h(\theta_0) + H^T(\theta_0)(\hat{\theta} - \theta_0) \\ &\approx h(\theta_0) + H^T(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \\ &= H^T(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \end{aligned}$$

For sufficiently large samples,

$$\begin{aligned} \sqrt{n}h(\hat{\theta}_n) &\approx H^T(\hat{\theta}_n)\sqrt{n}(\hat{\theta}_n - \theta_0) \\ &\sim Normal(0, H_{\theta_0}^T B_{\theta_0}^{-1} H_{\theta_0}) \end{aligned}$$

Taking the product of these normals, we find that

$$nh(\hat{\theta}_n)(H_{\theta_0}^T B_{\theta_0}^{-1} H_{\theta_0})^{-1} h(\hat{\theta}_n)^T \sim \chi_r^2$$

For sufficiently large samples, we may evaluate the matrices at $\hat{\theta}_n$ instead of θ_0. ∎

Theorem 11.5 Chi-squared test. *Under the same assumptions,*

$$n(D\log(f_{\tilde{\theta}_n}(x)))^T B_{\theta_0}^{-1} D\log(f_{\tilde{\theta}_n}(x)) \sim \chi_r^2$$

Proof Using a Taylor expansion about $\hat{\theta}_n$:

$$\begin{aligned} \sqrt{n}D\log(f_{\tilde{\theta}_n}(x)) &\approx \sqrt{n}D\log(f_{\hat{\theta}_n}(x)) + \sqrt{n}D^2\log(f_{\hat{\theta}_n}(x))(\tilde{\theta}_n - \hat{\theta}_n) \\ &= -D^2\log(f_{\hat{\theta}_n}(x))\sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) \\ &\approx nB_{\theta_0}(\hat{\theta}_n - \theta_0) \end{aligned}$$

Taking the square of this again gives a χ_r^2 distribution. ∎

Note that the Wald test is useful when the restricted MLE is hard to calculate, and the chi-square test is useful when the unrestricted MLE is hard to calculate.

If we are testing multiple hypothesis, we should be careful about whether we are controlling the size for each individual test or the overall size. (Controlling the size overall tends to be more correct, but makes the power worse.)

12 Bayesian Statistics

In Bayesian statistics, the uncertainty about the value of the parameter θ is expressed by treating θ as a random variable. That is, $\theta \sim \pi(\theta)$ before any data is collected. We call $\pi(\theta)$ the *prior density*; this represents all the information we have before new data is collected. Once we observe the data, $X \mid \theta \sim P_\theta(x)$, we may use Bayes' rule to compute a posterior probability:

$$\begin{aligned} \pi(\theta \mid x) &= \frac{P(x \mid \theta)\pi(\theta)}{\int P(x \mid \theta)\pi(\theta) d\theta} \\ &= \frac{P(x \mid \theta)\pi(\theta)}{f(x)} \\ &\propto P(x \mid \theta)\pi(\theta) \end{aligned}$$

(The last step follows because the denominator is a normalizing constant that does not depend on θ.) However, the choice of a prior can be challenging to justify.