STATISTICAL INFERENCE STS 412

LINEAR MODELS: GENERALIZED LINEAR MODELS

INTRODUCTION AND BASIC IDEAS

The simplest relationship between two variables is a linear one, namely

$$Y = \alpha + \beta X \quad (1)$$

where α and β are unknown parameters indicating the intercept and slope of the function. Other relationships between two variables include $Y = \alpha e^{\beta X}$, $Y = \alpha X^{\beta}$, $Y = \alpha + \beta \frac{1}{X}$. However, the first and second relations can be reduced to linear form in transformed variable by taking logs of both sides to give $log_e Y = log_e \alpha + \beta X$ and $log_e Y = log_e \alpha + \beta log_e X$. The third relation may be regarded as linear in the variables Y and $1/X$.

The Generalized Linear Model is a generalization of the general linear model (of which, for example, multiple regression is a special case. The generalized linear model differs from the general linear model in two major respects. Firstly, the distribution of the dependent variable (or response variable) can be non-normal, and does not have to be continuous, i.e. it can be binomial, multinomial or ordinal multinomial (i.e. contain information on ranks only). Secondly, the dependent variable values are predicted from a linear combination of predictor variables, which are connected to the dependent variable via a link function (to be explained).

SIMPLE LINEAR REGRESSION MODEL IN MATRIX NOTATION

By definition, the true regression of Y on X consists of the means of populations of Y values where a population is determined by the X value. The mathematical description of an observation is given by

$$Y_i = \mu_{Y.X} + \varepsilon_i \quad (1.1a)$$

$$= \alpha + \beta X_i + \varepsilon_i \quad (1.1b) \qquad (1.1)$$

$$= \mu + \beta(X_i - \bar{X}) + \varepsilon_i \quad (1.1c)$$

where α and β are parameters to be estimated and X is an observable parameter; α represents the population Y intercept, the value of μ_Y at $X = 0$, that is $\mu_{Y.0}$; β is the slope of the line through the means of the Ypopulation. In the last form of 1.1, that is 1.1c with the X's measured from their mean, μ is estimated by $\bar{Y}$. The ε's are assumed to be from a single population with zero mean and variance σ^2. This variance is another parameter to be estimated.

Writing the simple regression model as $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1,2,\ldots,n,$ we can write the set of observations as

$$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$$

$$\dots\dots\dots\dots\dots\dots\dots\dots$$

$$Y_n = \beta_0 + \beta_1 X_n + \varepsilon_n$$

In vector and matrix notation, we can write the above set of equations as

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_0 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Writing compactly we have:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_0 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or $\boxed{Y = \beta X + \varepsilon}$ or y= $Xb + e$

For the above model, we make the following assumptions presenting them in matrix form. For the assumption that the ε's have zero mean, we write $E(\varepsilon) = 0 =_n 0_1$, the null vector, *E* is for expectation or mean. Thus $Y = \beta X + \varepsilon \implies E(Y) = X\beta$. Note that

$$\varepsilon\varepsilon' = \begin{pmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 \varepsilon_1\varepsilon_3 & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 \; \varepsilon_2\varepsilon_3 & \varepsilon_2\varepsilon_n \\ \cdots & \cdots \quad \cdots & \cdots \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2\varepsilon_n\varepsilon_3 & \varepsilon_n^2 \end{pmatrix}$$

Squares on the diagonal suggest variances, off diagonal cross products suggest covariances. Consequently, assumptions concerning homogeneous variance and uncorrelated errors may be stated as follows:

$$E(\varepsilon\varepsilon') = \begin{pmatrix} \sigma^2 & 0 \; 0 \; \dots & 0 \\ 0 & \sigma^2 0 \; \dots & 0 \\ \cdots & \cdots \; \cdots & \cdots \\ 0 & 0 \;\; 0 \; \dots & \sigma^2 \end{pmatrix}$$

$$= I\sigma^2 \text{ or } I\sigma_{Y.X}^2$$

where*I* is the identity matrix.

REMARKS

1. From The sample, β is estimated by $\hat{\beta}$ or b.

2. To estimate the population mean when $X = X_0$, we write $X'_0 = (1 \quad X_0)$ and the sample regression equation is

$$\hat{\mu}_Y = X'_0\beta \text{ or } \hat{Y} = X'_0\hat{\beta}$$

3. $X'X$is the matrix of the model given by $X'X = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}_{2 \times 2}$

4. $X'Y$is a 2 x 1 where $X'Y = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}_{2 \times 1}$

5. If $X'X$ is nonsingular, its inverse $(X'X)^{-1} = \frac{1}{n\sum X_i^2 - (\sum X_i)^2}\begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{pmatrix}$

6. $\hat{\beta} = b = \begin{pmatrix} \bar{Y} - \hat{\beta}_1\bar{X} \\ \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n} \end{pmatrix}$

7. By definition, $e = Y - \hat{Y}$ is a deviation of the observation from the regression line. It is called a residual. We note that the sum of squares of the residual is a measure of the overall failure of the model to fit the data. We can write

$$\begin{aligned} e'e &= SS(\text{residual}) \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= Y'Y - \hat{\beta}'X'Y \end{aligned}$$

8. It will be shown that σ^2 can be estimated using the simple residual sum of squares.

9. $SS(\text{model}) = \hat{\beta}'X'Y$

10. According to Gauss Markov, the model $Y = \beta X + \varepsilon$ may also be denoted by $(Y, \beta X, \sigma^2 I_n)$.

ESTIMATION ($\boldsymbol{\beta}, \boldsymbol{\sigma^2}$) AND PROPERTIES

Basically, there are 4 methods of estimation which lead to the same estimator under certain frequently-used assumptions. All the four procedures are summarized in terms of the full rank model, where in $Y = X\beta + \varepsilon$, X has full column rank, $E(Y) = X\beta$ and $E(\varepsilon) = 0$.

1. **Ordinary Least Square (OLS)**

This involves choosing $\hat{\beta}$ or b as the value of β which minimizes the sum of squares of deviations of the observations from their expected value, i.e. choose $\hat{\beta}$ as that β which minimizes

$$Y = (\quad)\sum_{i=1}^{N}[y_i - E(Y_i)]^2 = (Y - X\beta)'(Y - X\beta)$$

The resulting estimator is, as we have seen $\hat{\beta} = (X'X)^{-1}X'Y = \hat{b}$ (To be proved).

2. **Generalized Least Squares**

On assuming that the variance-covariance matrix of *e* is *Var(e) = V*, this method involves minimizing $(Y - Xb)'V^{-1}(Y - Xb)$ with respect to b. This leads to

$$\overleftrightarrow{b} = (X'V^{-1}X)^{-1}X'V^{-1}Y \text{ or } \tilde{b} \text{ or } b^0$$

Remark: When $V = \sigma^2 I$, the generalized and the OLS estimators are the same: $\overleftrightarrow{b} = \hat{b}$

3. **Maximum Likelihood**

With least squares estimation no assumption is made about the form of the distribution of the random error terms in the model. On assuming that the errors, *e*'s are normally distributed with zero mean and variance-covariance matrix *V*, i.e. $e \sim N(0, V)$, the likelihood is

$$L = (2\pi)^{-\frac{1}{2}p}|V|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(Y - X\hat{\beta})'(V^{-1})(Y - X\hat{\beta})\right\}$$

Maximizing with respect to $\hat{\beta}$ is equivalent to solving $\delta(log_e L)/\delta\hat{\beta} = 0$ which gives $b^0 = (X'V^{-1}X)^{-1}X'V^{-1}Y$; the same as generalized least squares estimator.

4. **BLUE: The Best Linear Unbiased Estimator**

We shall take this up fully.

TWO REMARKS (ABOUT THESE ESTIMATORS)

1. Least squares estimation does not pre-suppose any distributional properties of the *e*'s other than finite zero means and finite variances.

2. M.I estimation under normality assumption leads to the same estimator, b^0, as generalized least squares; and this reduces to the ordinary least squares estimator b or $\hat{\beta}$ when $V = \sigma^2 I$.

Again, from equation $Y = \beta X + e$, $e = Y - X\beta$

$$e' = (Y - X\beta)'$$

Sum of squares of the observation, $SS = e'e$

$$SS = (Y - X\beta)'^{(Y-X\beta)} = e'e$$

$$= Y'Y - Y'(X\beta) - (X\beta)'Y + (X\beta)'X\beta$$

$$= Y'Y - \beta XY - \beta'X'Y + \beta'X'X\beta$$

$$= Y'Y - 2\beta'X'Y + \beta'X'X\beta \qquad (1.4)$$

To estimate β by OLS, we differentiate (1.4) with respect to β and set resultant effect to zero to obtain an estimate $\hat{\beta}$ or simply b.

$$\frac{\delta s}{\delta \hat{\beta}} = 0 \Longrightarrow 2X'Y$$

$$\frac{\delta s}{\delta \hat{\beta}} = -XY' + \beta'X'X = 0$$

$$\beta'X'X = XY' \qquad (1.5)$$

Premultiply (1.5) by $(X'X)^{-1}$, to get

$$\hat{\beta}(X'X)^{-1}X'X = \hat{\beta}(X'X)^{-1}X'Y, \text{ where } I = (X'X)^{-1}X'X$$

$$\hat{\beta}I = (X'X)^{-1}X'Y$$

$$\boxed{\hat{\beta} = (X'X)^{-1}X'Y} \qquad (1.6)$$

1.3 Variance or Dispersion of Parameters in a Functional Regression Model

Given $\hat{\beta} = (X'X)^{-1}X'Y$

$V(\hat{\beta}) = V[(X'X)^{-1}X'Y]$

$= V(M'Y)$ where $M' = (X'X)^{-1}X'$

$= M'V(Y)M \qquad M = X(X'X)^{-1}$

$= (X'X)^{-1}X'V(Y)X(X'X)^{-1}$ where $V(Y) = \sigma^2$

$$= (X'X)^{-1}X'\sigma^2X(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1}X'X(X'X)^{-1}$$

$$= \sigma^2 i(X'X)^{-1}$$

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1} \quad (1.7)$$

1.4 Properties of Least Square Estimator

One of the properties of a good estimator is unbiasedness. From a Gauss-Markov model $(Y, X\beta, \sigma^2 I_n)$ where the estimate of $\beta$is $\hat{\beta} = (X'X)^{-1}X'Y$. It can be investigated if $\hat{\beta}$ is unbiased of β, where it is required to show that $E(\hat{\beta}) = \beta$

$$E(\hat{\beta}) = E[(X'X)^{-1}X'Y] \quad \text{where } Y = X\beta$$

$$= E[(X'X)^{-1}X'X\beta]$$

$$= E[I\beta]$$

$$\boxed{E(\hat{\beta}) = \beta} \quad (1.8)$$

1.5 Estimation of Variance, σ^2

1.5.1 Residual Error Sum of Square

The residual vector can be stated as original – predicted value

$$Y - \hat{Y} = Y - E(Y) \qquad Y = X\hat{\beta}$$

$$Y - X\hat{\beta} \text{where } \hat{\beta} = (X'X)^{-1}X'Y, \text{ we have } Y - X(X'X)^{-1}X'Y$$

$$[I - X(X'X)^{-1}X']Y \quad (1.7)$$

by denoting $I - X(X'X)^{-1}X'$ by M, we have MY, M is idempotent symmetric matrix

$$M'M = MM = M \qquad MX = 0$$

$$[I - X(X'X)^{-1}X']X = 0$$

$$X - X(X'X)^{-1}X'X$$

$$X - X = 0 \quad \text{required}$$

$$SSE = (Y - \hat{Y})'(Y - \hat{Y})$$

$$= (MY)'(MY)$$

$$= Y'M'MY$$

$= Y'MY$

$$SSE = Y'[1 - X(X'X)^{-1}X']Y \tag{1.8}$$

$= Y'Y - Y'X(X'X)^{-1}X'Y$

$$= Y'Y - (X'X)^{-1}X'YX'Y = Y'Y - \hat{\beta}X'Y \tag{1.9}$$

1.5.2 Estimating σ^2

$SSE = Y'[I - X(X'X)^{-1}X']Y$ since $Y \sim N(X\beta, \sigma^2)$

If there exist $X \sim N(\mu, V)$, then

$$E(X'AX) = tr(AV) + \mu'A\mu$$

Proof:

Given $E(X) = \mu, \quad Var(X) = V$

$$V(X) = E[X - \mu][X - \mu]'$$

$= E[X'X - X\mu' - \mu X' + \mu\mu']$

$= E[X'X] - [X\mu'] - E[\mu X'] + E(\mu\mu')$

$= E[X'X] - E(X)\mu' - \mu E(X') + \mu\mu'$

$= E[X'X] - \mu\mu' - \mu\mu' + \mu\mu'$

$$= E[X'X] - \mu\mu' = \boxed{E(X'X) = V + \mu\mu'} \tag{1.10}$$

$E[X'AX] = E[tr(X'AX)]$

$$= E[tr(AXX')]$$

$$= tr[AE(XX')]$$

$$= tr[A(V + \mu\mu')]$$

$$= tr(AV) + tr(\mu'A\mu)$$

$$= tr(AV) + \mu'A\mu$$

Applying this theorem and noting that $Y = X$

$$A = [I - X(X'X)^{-1}X'], \quad \mu = X\beta, \quad V = \sigma^2 I$$

$$E(SSE) = tr[I - X(X'X)^{-1}X']\sigma^2 I + \beta'X'[I - X(X'X)^{-1}X']X\beta$$

PROOFS ON GLM's

Assumption:

A linear relationship between a variable Y and $k - 1$ explanatory variables $X_2, X_3 \ldots, X_k$ and a disturbance term *U*. If we have a sample of *n* observations on *Y* and the *X*'s we can write

$$Y_i = \beta_1 + \beta_{2i}X_{2i} + \beta_{3i}X_{3i} + \cdots + \beta_k X_{ki} + U_i, \quad i = 1,2,\ldots,n \tag{1}$$

or $$Y = X\beta + \mu \tag{2}$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & X_{21}\cdots & X_{k1} \\ 1 & X_{22}\cdots & X_{k2} \\ \vdots & & \vdots \\ 1 & X_{2n}\cdots & X_{kn} \end{pmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix} \tag{3}$$

$E(U) = 0$, $E(UU') = \sigma^2 I$, *X*has rank *k*<*n*, *X* is a set of fixed numbers.

$$E(UU') = \begin{bmatrix} E(U_1^2) & E(U_1U_2)\cdots & E(U_1U_n) \\ E(U_2U_1) & E(U_2^2)\ \cdots & E(U_2U_n) \\ \vdots & & \vdots \\ E(U_nU_1) & E(U_nU_2)\cdots & E(U_n^2) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0\cdots & 0 \\ 0 & \sigma^2\cdots & 0 \\ \vdots & & \vdots \\ 0 & 0\cdots & \sigma^2 \end{bmatrix} \tag{4}$$

PROOF OF LEAST SQUARES ESTIMATES

Let $\beta = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$ denote a column vector of estimates of β then

$$Y = X\hat{\beta} + e \tag{5}$$

Where *e*denotes the column vectors of *n* residuals $Y - X\hat{\beta}$

$$e'e = \sum_{i=1}^{n} e_i^2$$

$$= \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right)$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \qquad (6)$$

Since $\hat{\beta}'X'Y$ is a scalar and thus equal to its transpose $Y'X\hat{\beta}$.

To find the value of $\hat{\beta}$ which minimizes the sum of squared residuals, we differentiate (6)

$$\frac{\partial}{\partial\hat{\beta}}(e'e) = -2X'Y + 2X'X\hat{\beta}$$

Equating to 0 gives

$$X'X\hat{\beta} = X'Y$$

$$\Longrightarrow \hat{\beta} = (X'X)^{-1}X'Y \qquad (7)$$

This is the fundamental result for the least squares estimators. Alternatively, we can write

$$\frac{\partial}{\partial\hat{\beta}}(e'e) = -2YX' + 2\hat{\beta}'X'X \quad \text{which gives}$$

$$\hat{\beta}' = Y'X(X'X)^{-1}$$

Transposing both sides of this last result takes us back directly to the fundamental result given by (7).

MEAN AND VARIANCE OF $\hat{\beta}$

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad (i)$$

But $Y = X\beta + \varepsilon$ (ii)

Using (ii) in (i) by substituting for Y we obtain

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + \varepsilon)$$

$$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon$$

$$= \beta + (X'X)^{-1}X'\varepsilon$$

$$E\left(\hat{\beta}\right) = E[(X'X)^{-1}X'\varepsilon]$$

$$= E(\beta) + E[(X'X)^{-1}X'\varepsilon]$$

$$= E(\beta) + (X'X)^{-1}X'.E(\varepsilon)$$

$= E(\beta)$

$= \beta$ showing that a is an unbiased estimator of b

$V(\hat{\beta}) = V[(X'X)^{-1}X'Y]$from (i) above

Let $A' = (X'X)^{-1}X'$, then add $A = X(X'X)^{-1}$

$$V(\hat{\beta}) = V(A'Y)$$

$= A'V(Y)A$

$$= (X'X)^{-1}X'\sigma^2 X(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1}X'X'X(X'X)^{-1}$$

$$= \sigma^2 I(X'X)^{-1}$$

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

PROPERTIES OF A GENERALIZED INVERSE

In our study of generalized linear models and statistical inference, a matrix of the form *X'X* often occurs. Suppose we take our matrix *A* to be *X'X*. We now state four (4) useful properties of *X'X* as follows:

Theorem: When *G* is a generalized inverse of *X'X*, then

(i) *G'* is also a generalized inverse of $X'X$;

(ii) $XGX'X = X$ i.e. GX' is a generalized inverse of *X*;

(iii) XGX' is invariant to *G*

(iv) XGX' is symmetric, whether *G* is or not.

Proof:

(i) By definition, *G* satisfies

$$X'XGX'X = X'X$$

Transposing gives

$$X'XG'X'X = X'X \quad \text{proved.}$$

(ii)&(iii) will be proved later since we need a lemma to be able to prove this.

Proof of property iv now follows

(iv) Let*S* be a symmetric generalized inverse of $X'X$

Then XSX' is symmetric

But $XSX' = XaX'$ and therefore XGX' is symmetric. Hence the theorem is proved.

Lemma 1:

$PX'X = QX'X$ implies $PX' = QX'$

Using this lemma, we now prove (iii) i.e. XaX' is invariant to *G*.

Proof: Suppose *F* is some other generalized inverse, different from *G*.

Then (ii) gives $XGX'X = XFX'X$. Applying this lemma 1, we have $XGX' = XFX'$ showing that XGX' is the same for all g-inverse of $X'X$.

Another Procedure

In the last procedure just described and verified (or proved), we stated that the matrix *A* was partitioned such that its leading minor is of rank *r* where *r* is the rank of *A*. in this method, which is about to be described, there is no need for the non-singular minor of order *r* to be in the leading position. Hence, an algorithm for finding a generalized inverse of *A* is as follows:

(i) In *A*, of rank *r*, find any non-singular minor of order *r*. Call it *M*.

(ii) Invert *M* and transpose the inverse: $(M^{-1})'$

(iii) In *A* replace each element of *M* by the corresponding element of $(M^{-1})'$

(iv) Replace all other elements of *A* by zero

(v) Transpose the resulting matrix. The result is *G*, a generalized inverse of *A*.

REMARKS AND PROPERTIES

In general, the algorithm must be carried out as described in steps (i) to (v). One case where it can be simplified is when *A* is symmetric $(A = A')$. Then only principal minor of *A* is symmetric and the transposing in both (iii) and (v) can be ignored. The algorithm can then become as follows:-

(i) In *A*, of rank *r* and symmetric, find any non-singular principal minor of order *r*. Call it *M*.

(ii) Invert *M*

(iii) In *A* replace each element of *M* by the corresponding element of M^{-1}.

(iv) Replace all other elements of *A* by zero. The result is *G*, a generalized inverse of *A*.

We now state and prove where necessary the properties of a generalized inverse.

THE GENERAL LINEAR HYPOTHESIS FOR THE MODEL $(\boldsymbol{Y}, \boldsymbol{X\beta}, \boldsymbol{\sigma_n^2})$

Let us consider the general hypothesis given by

$$H: k'b = m \qquad (1)$$

where$b$ is the k + 1 order vector of parametes of the model;

$k'$is any matrix of s rows and k + 1 columns; and

m is a vector of order s of specified constants.

Remarks:

(a) One limitation of k' is that it must have full row rank i.e. $r(k') = s$

(b) Four hypotheses of particular interest of which equation (1) above is the general form are:

(i) $H: b = 0$, the hypothesis that all elements of b are zero.

(ii) $H: b = b_0$, the hypothesis that b_i and $b_{i0}$for $i = 0,1,2, \ldots k$, i.e. that each b_i is equal to some specified value b_{i0}

(iii) $H: \lambda' b = m$, that some linear combination of the elements of b equals a specified constant

(iv) $H: b_q = 0$, that some of the b_i's, q of them where $q<k$, are zero.

We now develop the F-statistic to test the hypothesis $H: k'b = m$. We already have the following:

$y \sim N(Xb, \sigma^2 I)$

$\hat{b} = (X'X)^{-1}X'y$

and$\hat{b} \sim N[b, (X'X)^{-1}\sigma^2]$

Therefore $k'\hat{b} - m \sim N[k'b - m, k'(X'X)^{-1}k\sigma^2]$.

Also it is true that when x is $N(\mu, v)$ then $z'Ax$ is $\chi^2\left[r(A), \frac{1}{2}'\mu' A\mu\right]$ if and only if AV is idempotent, the following quadratic, in $k'\hat{b} - m$, using $[k'(X'X)^{-1}k]^{-1}$ as the matrix of the quadratic, has a χ^2 distribution: if

$$Q = (k'\hat{b} - m)'[k'(X'X)^{-1}k]^{-1}(k'\hat{b} - m)$$

Hence, $Q/\sigma^2 \sim \chi^2\{s, (k'b - m)'[k'(X'X)^{-1}k]^{-1}(k'b - m)/2\sigma^2\}$ (2)

The independence of Q and SSE is not shown when $x \sim N(u, \mu)$ the quadratic forms $X'AX$ and $X'BX$ are distributed independently if and only if $AVB = 0$ (or, equivalently, $BVA = 0$). To do this, we first express Q and SSE as quadratic forms of the same normally distributed random variable, noting initially that the inverse of $k'(X'X)^{-1}k$ used in (2) exists because k' has full row rank and $X'X$ is symmetric. Then, on replacing $\hat{b}$by $(X'X)^{-1}X'y$, equation (2) for Q becomes

$$Q = [k'(X'X)^{-1}X'y - m]'[k'(X'X)^{-1}k]^{-1}[k'(X'X)^{-1}X'y - m]$$

Therefore,

$$k'(X'X)^{-1}X'y - m = k'k'(X'X)^{-1}X'[y - Xk(k'k)^{-1}m]^{-1}$$

and so

$$Q = [y - Xk(k'k)^{-1}m]'X(X'X)^{-1}k[k'(X'X)^{-1}k]^{-1} \times k'(X'X)^{-1}X'^{[y - Xk(k'k)^{-1}m]}.$$

Now consider the error sum of squares

$SSE = y'[1 - Xk'(X'X)^{-1}X']y.$

Because the products $X'[1 - X(X'X)^{-1}X']$ and $[1 - X(X'X)^{-1}X']X$ are both null, SSE can be written as

$$SSE = [y - Xk(k'k)^{-1}m]'[1 - X(X'X)^{-1}X'][y - Xk(k'k)^{-1}m].$$

Both Q and SSE have now been expressed as quadratics in the vector $y - Xk(k'k)^{-1}m$. Although we already know that Q/σ^2 and SSE/σ^2 have χ^2 distributions, this is further seen form their being quadratics in $y - Xk(k'k)^{-1}m$ which is normally distributed vector; and the matrix in each quadratic is idempotent. But, more importantly, the product of the two matrices is null

$$[1 - X(X'X)^{-1}X']X(X'X)^{-1}k[k'(X'X)^{-1}k]^{-1}k'(X'X)^{-1}X' = 0$$

Therefore, by the independence of Q and SSE as illustrated above, Q and SSE are distributed independently. Hence

$$F(H) = \frac{Q/s}{SSE/[N - r(X)]} = Q/S\sigma^2$$

$$\sim F'\{S, N - r(X), (k'b - m)'[k'(X'X)^{-1}k]^{-1}(k'b - m)/2\sigma^2\} \qquad (3)$$

and under the null hypothesis $H: k'b = m$

$$F(H) \sim F_s, N - r(X)$$

Hence $F(H)$ provides a test of the hypothesis $H: k'b = m$

Thus, the F-statistic for testing the hypothesis $H: k'b = m$ is

$$\boxed{F(H) = \frac{Q}{S\sigma^2} = \frac{(k'b-m)'\left[k'(X'X)^{-1}k\right]^{-1}(k'b-m)}{S\sigma^2}} \qquad (4)$$

with$s$ and $N - r$ degrees of freedom, s being the number of rows in k', it being of full row rank.

REMARKS/IMPORTANT STATEMENTS

(*) As long as k' has full row rank, $F(H) = \frac{Q}{S\sigma^2}$ given in (4) above and enclosed for emphasis can be used to test any linear hypothesis whatever. Nomatter what hypothesis is, it has only to be written in the form $k'b = m$ and $F(H)$ of equation (4) above provides the test.

NONPARAMETRIC OR DISTRIBUTION FREE STATISTICS

Introduction

What are nonparametric statistical procedures? Thus far, the statistical procedures discussed or used have some underlying assumptions for which the procedures are valid. These techniques are for the estimation of parameters and for testing the hypothesis concerning them. The assumptions generally specify the form of the distribution and in most cases are concerned largely with data where the underlying distributions is normal. ARE MOST VARIABLES NORMALLY DISTRIBUTED? OF COURSE NO.

However, a considerable amount of data of interest is such that the underlying distribution is not normal and is not specified. For instance, environmental data are highly skewed. Apart from normality condition, another factor that often limits the applicability of test based on the assumption that the sampling distribution is normal is the size of the sample available for the analysis. If our sample is very small, parametric procedures cannot be used because there is no way to test the assumption of normality.

Another factor that often limits the applicability of test that are based on the normality assumptions is lack of precise measurement (problems in measurement). This general measurement issue has to do with types of measurement or scale of measurement. Data in form of counts, ranks, or the signs of differences for paired observations do not meet the assumptions of normality, therefore cannot be analyzed using parametric procedures.

What then are Nonparametric or Distribution free statistics? Nonparametric statistics are statistics where it is not assumed that the population fits any parametrised distribution. Nonparametric statistics are typically applied to population that takes on a ranked order. Nonparametric procedures differ from parametric models because the distribution is not specified apriori but it is instead determined from the data. The term nonparametric is not meant to imply that such models completely lacked parameters but that the number and the nature of parameters are flexible and not fixed in advanced. Hence nonparametric models are also called distribution-free or parameter free.

Brief Overview of Nonparametric Methods

Basically nonparametric or distribution free inferential statistical methods are mathematical procedures for statistical hypothesis testing which, unlike parametric statistics makes no assumptions about the frequency distribution of the variables being accessed. Some of the most frequently nonparametric tests used include:

1. Binomial test
2. Anderson-Darling test
3. Chi-square test
4. Cochran's Q
5. Coher's kappa 6.
6. Efron-Petrosian Test
7. Fisher's exact test
8. Friedman two-way analysis of variance
9. Kendall's tau
10. Kendall's W
11. Kolmogorov-Smirnov Test
12. Kruskal Wallis one-way analysis of variance by ranks
13. Kniper's test
14. Mann-Whitney U or Wilcoxon rank sum test
15. McNemar's test (a special case of the chi-squared test)
16. median Test
17. Pitman's permutation test
18. Siegel-Turkey test
19. Spearman's rank correlation coefficient
20. Student-Newman-Keuls (SNK) test
21. Wald-Wolfowitz runs test
22. Wilcoxon signed-rank test

There is at least one nonparametric equivalent for each parametric general type of test. In general, these tests fall into the following categories:

(i) Tests of differences between groups (independent samples)
(ii) Tests of differences between variables (dependent samples)
(iii) Tests of relationships between variables.

Differences between independent groups – usually, when we have 2 samples that we want to compare concerning their mean value for some variable of interest, we will use the *t*-test for independent samples; nonparametric alternatives for this test are Wald-Wolfoyitz runs test, the Mann-Whitney U test, and the Kolmogorov-Smirnov two sample test. If we have multiple groups, we would use analysis of variance; the nonparametric equivalents to this method are the Kruskal-Wallis analysis of ranks and the Median test.

Differences between dependent groups – If we want to compare two variables measured in the same sample, we would customarily use the *t*-test for dependent samples (in Basic Statistics for example, if we wanted to compare students' math skills at the beginning of the semester with their skills at the end of the semester). Nonparametric alternatives to this test are the sign test and Wilcoxon's matched pairs test. If the variables of interest are dichotomous in nature (i.e., "on", vs "off", "pass" vs "no pass") then McNemar's chi-square test is appropriate. If there are more than two variables that were measured in the same sample, then we would customarily use repeated measures ANOVA. Nonparametric alternatives to this method are Friedman's two-way analysis of variance and Cochran Q test (if the variable was measured in terms of categories e.g. "passed" vs "failed"). Cochran Q is particularly useful for measuring changes in frequencies (proportions) across time.

Relationships between variables – To express a relationship between two variables, one usually computes the correlation coefficient. Nonparametric equivalents to the standard correlation coefficient are Spearman R, Kendall Tau, and Coefficient Gamma. If the two variables of interest are categorical in nature (e.g. "passed" vs "failed" by "male" vs "female"), appropriate nonparametric statistics for testing the relationship between the two variables are the chi-square test, the Phi coefficient, and the Fisher exact test. In addition, a simultaneous test for relationships between multiple cases is available: Kendall coefficient of concordance. This test is often used for expressing inter-rater agreement among independent judges who are rating (ranking) the same stimuli.

THE KOLMOGOROV-SMIRNOV TEST (K-S TEST)

Introduction

The K-S test is a test of goodness of fit test. K-S finds out (or tests) whether two independent samples have been drawn from populations having the same cumulative distribution. There are 1-sampl;e and 2-sample K-S tests procedures;

1-Sample K-S Test: This is concerned with the agreement between the distribution of a set of sample values and some specified theoretical distribution.

2-Sample K-S Test: The 2-sample K-S test is concerned with the agreement between two cumulative distributions (cumulative relative frequency distributions). That is agreement between two sets of sample values.

DESCRIPTION/PROCEDURE

To apply K-S two-sample test, we proceed as follows:

1. Arrange each of the two groups of scores in a cumulative frequency distribution by using the same intervals for both groups.

2. Denote $F_{n_1}(x)$ the observed cumulative distribution for one sample of size n_1 where $F_{n_1}(x) = k/n_1$, where k is the number of data equal to or less than x. Similarly, let $F_{n_2}(x) = k/n_2$ denote the observed cumulative distribution of the second sample of size n_2.

3. Determine the largest of these differences. Call it

$$D_{n_1 n_2} = \underset{\text{supremum}}{\text{maximum}} \left| F_{n_1}(x) - F_{n_2}(x) \right| \qquad (1.1)$$

K-S Statistic

The Kolmogorov-Smirnov statistic is given by

$$D_n = \sup_x |F_n(x) - F(x)| \qquad (1.2)$$

Test Criteria

The method of determining the significance of the observed $D_{n_1 n_2}$ depends on the sample sizes n_1 and n_2 and the nature of the alternative hypothesis H_a or H_1.

(a) When n_1 and n_2 are both less than or equal to 25, use the appropriate table for a two-tailed test.

(b) For a two-tailed test when n_1 or $n_1 > 25$, use the appropriate table.

(c) For a one-tailed test when either n_1 or n_2 is >25, the value of χ^2 is computed from $\chi^2 = 4D_{n_1 n_2}^2 \cdot \frac{n_1 n_2}{n_1 + n_2} \sim \chi^2$ with df = 2.

NONPARAMETRIC CORRELATION

We consider only the following measures of association among variables. These are

1. The Spearman Rank Order Correlation Coefficient, r_s

2. The Kendall Coefficient of Concordance, W.

Formulae:

A. $r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$ where the symbols have their usual meanings.

Proof: It is left as an exercise to be $t = r_s\sqrt{\frac{n-2}{1-r_s^2}} \sim$ as studentst with $n - 2$ *df*.

B. **Test Statistics**

Kendall (1948) showed that for fairly large *n*, $n > 8$.

$t_{n-2} = \sqrt{\frac{(n-2)r_s^2}{1-r_s^2}}$ is distributed as student *t* with degree of freedom, $df = n - 2$.

REMARK

Use the last statistic $Z = r_s\sqrt{n-1}$ and compare results.

1. By the Glivenko-Cantelli theorem, if the sample comes from distribution $F(x)$, then D_n converges to 0 almost surely.

2. The Kolmogorov distribution is the distribution of the random variable $K = \sup_{t\in[0,1]} |B(t)|$ where $B(t)$ is the Brownian bridge. The cumulative distribution of *K* is given by

$$P_r(K \le x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty} e^{-(2i-1)^2\pi^2/(8x^2)}$$

3. **K-S Test**

3.1 Under null hypothesis that the sample comes from the hypothesized distribution $F(x)$,

$\sqrt{n}D_n \overset{n\to\infty}{\longrightarrow} \sup_t |BF(t)|$ in distribution, where $B(t)$ is the Brownian bridge.

3.2 If F is continuous, then $\sqrt{n}D_n$ converges to the Kolmogorov distribution which does not depend on *F*.

3.3 The goodness-of-fit test or the K-S test is constructed by using the critical values of the Kolmogorov distribution.

The null hypothesis is rejected at level α if $\sqrt{n}D_n > K_\alpha$ where K_α is found from $P_r(K \le K_\alpha) = 1 - \alpha$.

3.4 **Remark:**

In the K-S statistic, the empirical distribution function F_n for i.i. d (independent and identically distributed) observations X_i is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i} \leq x$$

where$I_{X_i} \leq x$ is the indicator function.

1. **Kendall Coefficient of Concordance, *W***

Kendal coefficient of concordance, *W*, measures the extent of association among several *k* sets of rankings of *N* entities. It is useful in determining the agreement among several judges or the association among three or more variables. It has special applications in providing a standard method of ordering entities according to consensus when there is available no objective order of the entities.

2. **Summary of Procedure**

These are the steps in the sue of *W*, the Kendall coefficient of concordance.

(1) Let *N* = the number of entities to be ranked, and
Let *k* = the number of judges assigning ranks
Last the observed ranks in a *K* x *N* table.

(2) For each entity, determine R_j, the sum of the ranks assigned to that entity by the *K* judges.

(3) Determine the mean of the *Rj*. Express each R_j as a deviation from the mean. Square these deviations, and sum the squares to obtain *s*. Note that

$$s = \sum\left(R_j - \frac{\sum R_j}{N}\right)^2 \text{ and that}$$

$$W = \frac{s}{\frac{1}{12}K^2(N^3 - N)}, \quad 0 < W < 1.$$

3. **Significance/Interpretation of *W***

The method of determining whether the observed value of *W* is significantly different from zero depends on the size of *N*:

(a) If *N* is 7 or smaller, table *R* gives critical values of *s* associated with *W*'s significant at the .05 and .01 levels.

(b) If *N* is larger than 7, use any of the following formulae:

(i) $\chi^2 = \frac{s}{\frac{1}{12}KN(N+1)}$with*af* = *N* – 1.

(ii) $\chi^2 = K(N-1)W$