Lecture Notes
Stat 4382
Intermediate Statistical Methods

J. D. Tubbs
Department of Statistical Science

Spring 2023

Contents

Preface

These notes have been modified from a sequence of statistical methodological courses that I usually teach to first year graduate statistics students at Baylor University. I have been using notes of this type for many years. These lecture notes will serve as the text for the course and a guide for the material that I will cover in my lectures.

In March 2020, Covid-19 arrived and our world changed! ZOOM and on-line course preparation became words that most of the seasoned faculty thought that they would never hear, let alone need and use. With this new reality what should my lecture notes contain, how do I present the new material, the software and the nuances of the statistical analysis in a ZOOM Meeting? You can teach an old dog new tricks but you can't make an old dog new! So how will the notes and the lectures change? The written notes will be pretty much as before, static presentation of the material that you need to learn and use. However, rather than provide examples of statistical analysis as static examples, I will use video materials in which I present, discuss, and comment on the additional examples that I use to supplement the notes and written material.

These videos will be available on canvas, they will be brief (10-20 minutes) with a single objective. The complete analysis of an example may take several videos. You will be able to view these at your convenience.

In addition, I have recently transitioned to create modules from the written material in my notes along with R and SAS code and output for specific examples using small to medium sized data sets from a variety of applications

The material that I have selected is based upon several assumptions. These include:

- You have had statistical methods courses in which you covered topics like hypothesis testing, confidence intervals, and regression models.
- You have taken some courses in mathematical statistics.
- You have had some experience with statistical software, such as, JMP, R, and perhaps SAS.
- You have had some experience with matrices and elementary linear algebra.

I often read and hear about how much the students and their parents want to have in person classes as in the days before the pandemic. In case you too long for the good ole days of in class lectures, consider the following

Figure 1: Remembering the Good Ole Days with in class Lectures

Part I

Descriptive and Inferential Methods

Chapter 1

One Population Methods

In this chapter we consider statistical methods that you may have already had. Review the material in the Appendix in order to familize yourself with the notation used in the notes. Suppose we have a single random variable, X, with a simple random sample of size n, given by $x_1, \ldots, x_n$. The objective is to "learn" about the random variable from the random sample where learning may include tables, graphs, and sample statistics. The specific type of tables, graphs, and statistics depend upon the type of data that is generated by X. That is, whether the random variable is discrete or continuous. Although the methods used are dependent upon the data type, they provide information concerning the location or middle of the data, the dispersion or spread of the data about the center, the shape of the data, and presence of unusual or questionable data within the sample (outliers or multi-modality).

The types of data are:

- Discrete
  - Nominal - {Male, Female}
  - Ordinal - {Freshmen, Sophomore, Junior, Senior}
  - Numerical - counts - cases of rare disease at Baylor University
- Continuous - Age, BMI, LDL

1.1 Discrete Variables

When the data are discrete, tables, graphs, and statistics are mainly counts or functions of counts, such as, proportions or percentages. There is not any useful shape information or outliers. There may be outcomes with small or large counts but that is what we can determine and illustrate with the tables, graphs, and sample statistics

Example: SAS output for a Discrete variable

The SAS code for the displays found in Figure 1.1 is

```
    options center nodate pagesize=100 ls=80;
*ods pdf;* style=journal;
libname consumer "/folders/myfolders/Large Data Sets/SAS Data Sets";
ods graphics on;
title1 'Kolache Sales in Large Consumer Warehouses';
```

```
data kolache; set consumer.kolache; texas = (state eq 'TX');
*  proc contents short data=kolache; run;

proc freq data=kolache; table state; run;
proc sgplot data=kolache;
vbar state;
run;
proc sgpie data=kolache;
pie state;
run;
```

1.2 Continuous Variables

In this section, we consider a continuous random variable, X, with its associated CDF, $F_X(x;\theta)$, and pdf, $f_X(X;\theta)$. Our objective is to use a random sample of size n for providing graphical methods and numerical descriptors for these functions and the parameters, θ.

1.2.1 Descriptive Statistics

The descriptive methods in this section consist of sample statistics and graphical displays. Both are useful and should be our first approach when we have data. The descriptive methods reveal characteristics about;

- the center or location of the random variable X,
- the dispersion or spread of the data (usually about the center),
- characteristics of the data that reveal the symmetry (or lack there of) and shape of the CDF or pdf. These could include values that appear to be suspect relative to the data one would expect to see, outliers or more than a single mode

Suppose that one has a random sample of size n from a CDF, $F_X(x;\theta)$, denoted by $x_1, \ldots, x_n$. The objective of this chapter is to defined statistics and sampling methods for $\theta \in \Theta$.

Measures of Location or Center

Mean

The weighted sample mean is calculated as

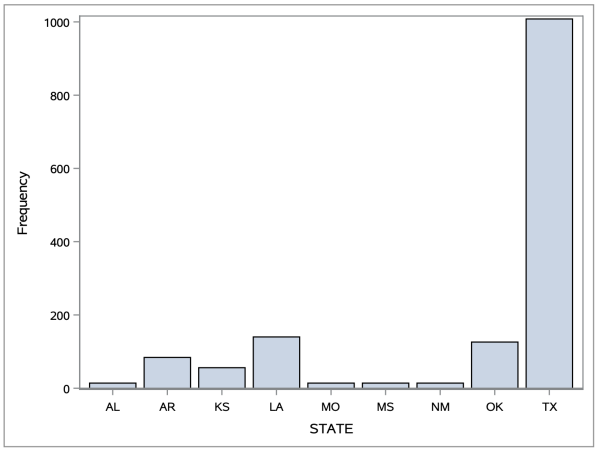
$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where n is the number of nonmissing values for a variable, x_i is the i^{th} value of the variable, and w_i is the weight associated with the i^{th} value of the variable. In the case where there is no WEIGHT variable, $w_i = 1/n$, one has the usual expression for $\bar{x}$ given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The FREQ Procedure

STATE				
STATE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AL	14	0.95	14	0.95
AR	84	5.71	98	6.67
KS	56	3.81	154	10.48
LA	140	9.52	294	20.00
MO	14	0.95	308	20.95
MS	14	0.95	322	21.90
NM	14	0.95	336	22.86
OK	126	8.57	462	31.43
TX	1008	68.57	1470	100.00



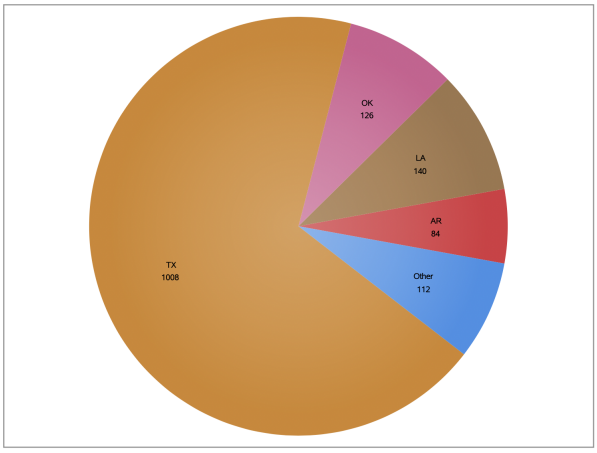


Figure 1.1: Discrete Kolache Data Displays

Median

The median is the value which divides the observations into two equal parts such that at least 50 % of the values are greater than or equal to the median and at least 50 % of the values are less than or equal to the median. The median is denoted by $\tilde{x}_{0.5}$. Suppose the sample of n observation are order (order statistics) as $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. The calculation of the median depends on whether the number of observations n is odd or even. When n is odd, then $\tilde{x}_{0.5}$ is the middle ordered value. When n is odd, $\tilde{x}_{0.5}$ is the the arithmetic mean of ($x_{(n/2)}$ and $x_{(n/2+1))}$.

Other measures for location include the ordered statistics which are used to estimate the quantiles and percentiles. If one suspects that there are outliers in the data set (either too small or too large) then trimmed or winsorized means can be computed.

Measures of Dispersion or Spread

Variance

The weighted sample variance is calculated as

$$\frac{1}{d}\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)^2$$

where n is the number of nonmissing values for a variable, x_i is the i^{th} value of the variable, $\bar{x}_w$ is the weighted mean, w_i is the weight associated with the i^{th} value of the variable, and d is the divisor controlled by the VARDEF= option in the PROC UNIVARIATE statement:

$$d = \begin{cases} n-1 & \text{if VARDEF=DF (default)} \\ n & \text{if VARDEF=N} \\ (\sum_i w_i) - 1 & \text{if VARDEF=WDF} \\ \sum_i w_i & \text{if VARDEF=WEIGHT — WGT} \end{cases}$$

If there is no WEIGHT variable, the formula reduces to

$$\frac{1}{d}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where $d = n$ or $d = (n-1)$.

Standard Deviation

The standard deviation is calculated as

$$s_w = \sqrt{\frac{1}{d}\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)^2}$$

or

$$s = \sqrt{\frac{1}{d}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Range and Interquartile Range

Consider a variable X with n observations $x_1, x_2, \ldots, x_n$. Order these n observations as $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. The range is a measure of dispersion defined as the difference between the maximum and minimum value of the data as $R = x_{(n)} - x_{(1)}$. The interquartile range is defined as the difference between the 75th and 25th quartiles as it covers the center of the distribution and contains 50 % of the observations.

Absolute Deviation

Another measure of dispersion is the "absolute median deviation" given by

$$D(\tilde{x}_{0.5}) = \frac{1}{n}\sum_{i=1}^{n} \mid x_i - \tilde{x}_{0.5} \mid .$$

Measures of Shape

Skewness

The sample skewness, which measures the tendency of the deviations to be larger in one direction than in the other, is calculated as follows:

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n} w_i^{3/2}\left(\frac{x_i - \bar{x}_w}{s_w}\right)^3$$

or

$$\frac{1}{n}\sum_{i=1}^{n} w_i^{3/2}\left(\frac{x_i - \bar{x}_w}{s_w}\right)^3$$

where n is the number of nonmissing values for a variable, x_i is the i^{th} value of the variable, $\bar{x}_w$ is the sample average, s is the sample standard deviation, and w_i is the weight associated with the i^{th} value of the variable. The sample skewness can be positive or negative; it measures the asymmetry of the data distribution and estimates the theoretical skewness, $\sqrt{\beta_1} = \mu_3 \mu_2^{-3/2}$ where μ_2 and μ_3 are the second and third central moments. Observations that are normally distributed should have a skewness near zero.

Kurtosis

The sample kurtosis, which measures the heaviness of tails, is calculated as follows:

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n} w_i^2\left(\frac{x_i - \bar{x}_w}{s_w}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

or

$$\frac{1}{n}\sum_{i=1}^{n} w_i^2\left(\frac{x_i - \bar{x}_w}{s_w}\right)^4 - 3$$

The sample kurtosis measures the heaviness of the tails of the data distribution. It estimates the adjusted theoretical kurtosis denoted as $\beta_2 - 3$, where $\beta_2 = \frac{\mu_4}{{\mu_2}^2}$, and μ_4 is the fourth central moment. Observations that are normally distributed should have a kurtosis near zero.[1]

[1] Note: other books might indicate the kurtosis of a normal variable as being 3. SAS subtracts 3 from the usual computation.

Coefficient of Variation (CV)

The coefficient of variation is calculated as a percent and is given by

$$CV = \frac{100 \times s_w}{\bar{x}_w}$$

or

$$CV = \frac{100 \times s}{\bar{x}}$$

Example (continued)

Figure 1.2 contains the descriptive statistics and extreme observations. Figure 1.3 contains the moments with trimmed and winsorized means.

The UNIVARIATE Procedure
Variable: unit_tot

Basic Statistical Measures			
Location		**Variability**	
Mean	38.77083	**Std Deviation**	17.06422
Median	38.50000	**Variance**	291.18775
Mode	41.00000	**Range**	94.00000
		Interquartile Range	23.00000

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0	266	94	1235
0	265	94	1236
0	264	94	1237
0	263	94	1238
0	262	94	1239

Figure 1.2: Basic Statistical Summary

The UNIVARIATE Procedure
Variable: unit_tot

Moments			
N	1008	**Sum Weights**	1008
Mean	38.7708333	**Sum Observations**	39081
Std Deviation	17.0642242	**Variance**	291.187748
Skewness	0.41848734	**Kurtosis**	0.34429592
Uncorrected SS	1808429	**Corrected SS**	293226.062
Coeff Variation	44.0130447	**Std Error Mean**	0.53747254

Trimmed Means								
Percent Trimmed in Tail	**Number Trimmed in Tail**	**Trimmed Mean**	**Std Error Trimmed Mean**	**95% Confidence Limits**		**DF**	**t for H0: Mu0=0.00**	**Pr > |t|**
5.06	51	38.28146	0.530991	37.23934	39.32357	905	72.09439	<.0001

Winsorized Means								
Percent Winsorized in Tail	**Number Winsorized in Tail**	**Winsorized Mean**	**Std Error Winsorized Mean**	**95% Confidence Limits**		**DF**	**t for H0: Mu0=0.00**	**Pr > |t|**
5.06	51	38.70833	0.531020	37.66616	39.75051	905	72.89424	<.0001

Figure 1.3: Additional Statistical Summaries

1.3 Useful Methods

In the next few sections I have included some topics that may be new to you. I have included them as an introductory topic.

1.3.1 Goodness-of-Fit Tests

In this section the test of hypothesis concerns the CDF for the random variable X rather than its parameters; (location and scale). The test can be stated as

$$H_0 : F_X(x) = F_{X_0}(x)$$

when $F_{X_0}(x)$ is completely specified.[2]
SAS provides the following goodness of fit tests[3] :

- Shapiro-Wilk
- Tests based upon the empirical CDF
  - Kolmogorov-Smirnov
  - Anderson-Darling
  - Cramer-von Mises

1.3.2 QQ and PP Plots

Probability plots can be used to assess normality of data, especially normality of the residuals in linear regression. Let $z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)}$ represent the ordered values of n independent and identically distributed $N(0,1)$ random variables. It can be shown that the expected value of $z_{(i)}$ is

$$E(z_{(i)}) \approx \gamma_i = \Phi^{-1}[(i - 3/8)/(n + 1/4)]$$

where Φ is the cdf for the standard normal given by

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-1/2t^2} dt.$$

The QQ plot consists of plotting of the ordered data (standardized residuals), $z_{(i)}$ versus γ_i, i.e. $(z_{(i)}, \gamma_i)$. If the data are normal then the resulting scatterplot should fall on the diagonal degree line $(z_{(i)} = \gamma_i)$. The PP plot is obtained when plotting of the ordered pairs $(\Phi(z_{(i)}), [i/n])$.

[2]When you specify the NORMAL option in the PROC UNIVARIATE statement or you request a fitted parametric distribution in the HISTOGRAM statement, the procedure computes goodness-of-fit tests for the null hypothesis that the values of the analysis variable are a random sample from the specified theoretical distribution.

[3]If you want to test the normality assumptions for analysis of variance methods, beware of using a statistical test for normality alone. A test's ability to reject the null hypothesis (known as the power of the test) increases with the sample size. As the sample size becomes larger, increasingly smaller departures from normality can be detected. Because small deviations from normality do not severely affect the validity of analysis of variance tests, it is important to examine other statistics and plots to make a final assessment of normality. The skewness and kurtosis measures and the plots that are provided by the PLOTS option, the HISTOGRAM statement, the PROBPLOT statement, and the QQPLOT statement can be very helpful. For small sample sizes, power is low for detecting larger departures from normality that may be important. To increase the test's ability to detect such deviations, you may want to declare significance at higher levels, such as 0.15 or 0.20, rather than the often-used 0.05 level. Again, consulting plots and additional statistics can help you assess the severity of the deviations from normality.

Shapiro-Wilk Statistic

The Shapiro-Wilk statistic, W (also denoted as W_n to emphasize its dependence on the sample size n) is the ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance (Shapiro and Wilk, 1965). The statistic W is always greater than zero and less than or equal to one ($0 < W \leq 1$). When the data are normally distributed, one would expect the ratio of these two estimators for the variance to be close to one, in which case, small values of W lead to the rejection of the null hypothesis of normality. The distribution of W is highly skewed. Seemingly large values of W (such as 0.90) may be considered small and lead you to reject the null hypothesis. The method for computing the p-value (the probability of obtaining a W statistic less than or equal to the observed value) depends on n. For $n = 3$, the probability distribution of W is known and is used to determine the p-value. For $n > 4$, a normalizing transformation is computed:

$$Z_n = \begin{cases} (-\log(\gamma - \log(1 - W_n)) - \mu)/\sigma & \text{if } 4 \leq n \leq 11 \\ (\log(1 - W_n) - \mu)/\sigma & \text{if } n \geq 12 \end{cases}$$

The values of σ, γ, and μ are functions of n obtained from simulation results. Large values of Z_n indicate departure from normality, and because the statistic Z_n has an approximately standard normal distribution, this distribution is used to determine the p-values for $n > 4$.[4]

EDF Goodness-of-Fit Tests

The EDF tests offer advantages over traditional chi-square goodness-of-fit test, including improved power and invariance with respect to the histogram midpoints. For a thorough discussion, refer to D'Agostino and Stephens (1986).

Definition: 1.1 *The* **Empirical Distribution Function (EDF)** *is defined for a set of n independent observations $X_1, \ldots, X_n$ with a common distribution function $F(x)$. Denote the observations ordered from smallest to largest as $X_{(1)}, \ldots, X_{(n)}$. The EDF, $F_n(x)$, is,*

$$\begin{array}{ll} F_n(x) = 0, & x < X_{(1)} \\ F_n(x) = \frac{i}{n}, & X_{(i)} \leq x < X_{(i+1)} \quad i = 1, \ldots, n-1 \\ F_n(x) = 1, & X_{(n)} \leq x \end{array} .$$

Note: $F_n(x)$ is a step function that takes a step of height $\frac{1}{n}$ at each observation. This function estimates the distribution function $F(x)$. At any value x, $F_n(x)$ is the proportion of observations less than or equal to x, while $F(x)$ is the probability of an observation less than or equal to x. EDF statistics measure the discrepancy between $F_n(x)$ and $F(x)$.[5]

Kolmogorov D Statistic

The Kolmogorov-Smirnov statistic (D) is defined as

$$D = \sup_x |F_n(x) - F(x)|$$

The Kolmogorov-Smirnov statistic is computed as the maximum of D^+ and D^-, where D^+ is the largest vertical distance between the EDF and the distribution function when the EDF is greater than the distribution

[4] The Shapiro-Wilks procedure is based upon "moment-type" estimators and would not be as powerful as the procedures that are based upon the estimated CDF for X. It should probably not be used without considering the other procedures.

[5] The computational formulas for the EDF statistics are based upon the probability integral transformation $U = F(X)$. That is, if $X \sim F(X)$ then $U \sim U(0, 1)$. In the test of fit problems, $F(X)$ is the null (or specified) distribution function.

function, and D^- is the largest vertical distance when the EDF is less than the distribution function[6].

$$\begin{array}{rcl} D^+ & = & \max_i \left(\frac{i}{n} - U_{(i)}\right) \\ D^- & = & \max_i \left(U_{(i)} - \frac{i-1}{n}\right) \\ D & = & \max\left(D^+, D^-\right) \end{array}$$

Quadratic EDF Statistics

The Anderson-Darling statistic and the Cramer-von Mises statistic are special cases of the general quadratic class of EDF statistics. This class of statistics is based on the squared difference $(F_n(x) - F(x))^2$. The general form of the quadratic class of EDF statistics is

$$Q = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \psi(x) dF(x) \tag{1.1}$$

where $\psi(x)$ is a weight function defined on the squared difference $(F_n(x) - F(x))^2$.

Anderson-Darling Statistic

The Anderson-Darling statistic considers $\psi(x) = [F(x)\,(1 - F(x))]^{-1}$ in which case equation (1.1) is

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i-1) \log U_{(i)} + (2n + 1 - 2i) \log(1 - U_{(i)})\right].$$

where $U_{(i)} = F_X(X_{(i)})$.

Cramer-von Mises Statistic

The Cramer-von Mises statistic considers $\psi(x) = 1$ in which case equation (1.1) is

$$W^2 = \sum_{i=1}^{n} \left(U_{(i)} - \frac{2i-1}{2n}\right)^2 + \frac{1}{12n}.$$

Example (continue)

Figure 1.4 has the result for a goodness of fit results, histogram with vertical boxplot and a QQ plot.

1.3.3 Kernel Density Estimation

In this section, the emphasis has changed, in that, we are interested in determining or estimating the shape of the pdf for the random variable X. One could use graphical methods, such as, histograms and stem-leaf plots. Instead, a computational method is used.

A procedure used for estimating a probability density function using the observed data is considered. As with histograms, the procedure is used to provide a graphical representation of the pdf of X using the observed data, $x_1, x_2, \ldots, x_n$. The estimate is called the *kernel density estimate*. Kernel density estimation is a nonparametric technique for density estimation in which a known density function (the kernel) is averaged across the observed data points to create a smooth approximation for $f_X(x)$. PROC KDE uses a Gaussian density as the kernel, and its assumed variance determines the smoothness of the resulting estimate. See Silverman (1986) for a thorough review and discussion.

[6]PROC UNIVARIATE uses a modified Kolmogorov D statistic to test the data against a normal distribution with mean and variance equal to the sample mean and variance.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.983172	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.055999	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.527241	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	3.93493	**Pr > A-Sq**	<0.0050

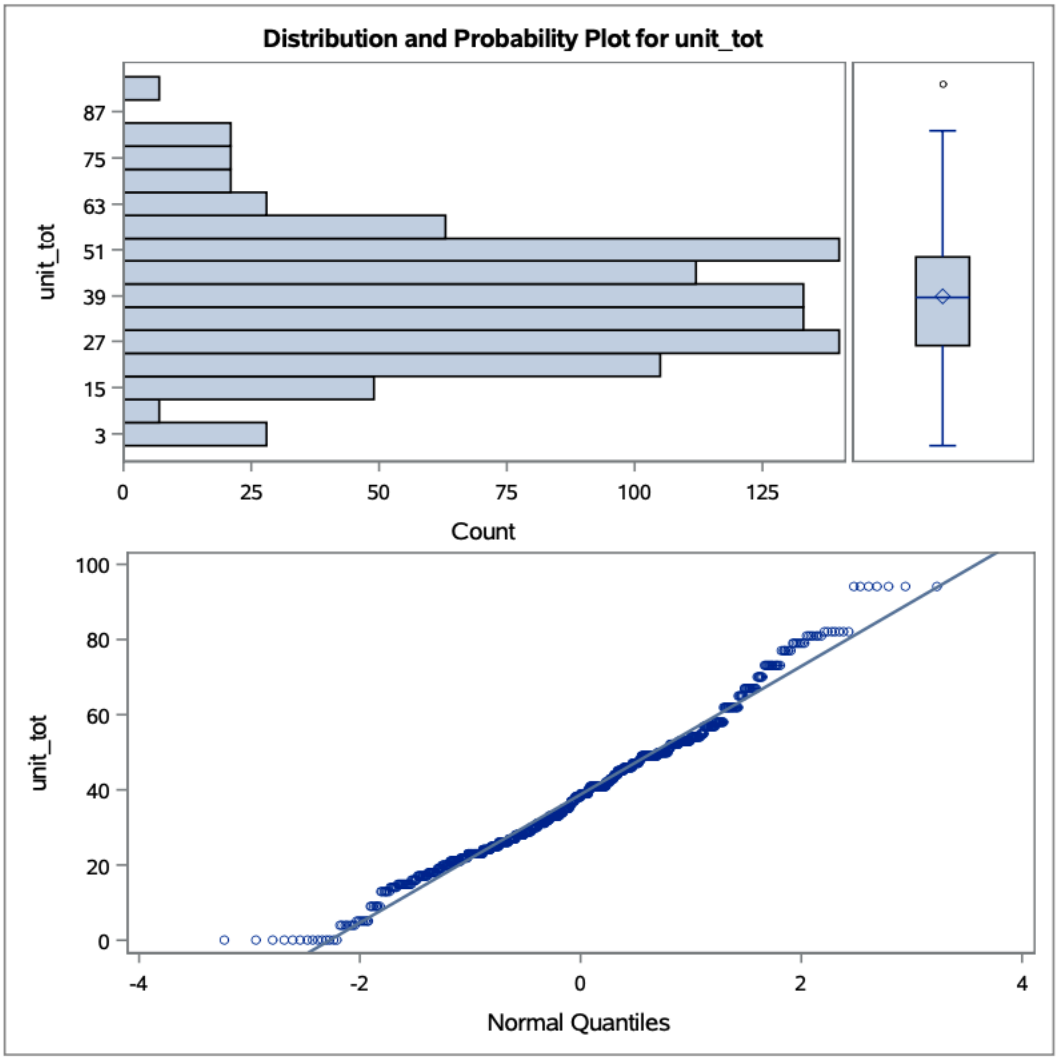


Figure 1.4: Additional Statistical Summaries

Computational Methods

Univariate Kernel Density Estimates - SAS

Let (X_i, W_i), denote the observed sample of X_i with specified weight W_i for $i = 1, 2, \ldots, n$. The weighted kernel density estimate of $f(x)$, the density of X, is

$$\hat{f}(x) = \frac{1}{\sum_{i=1}^{n} W_i} \sum_{i=1}^{n} \varphi_h(x - X_i)$$

where h is the bandwidth and

$$\varphi_h(x) = \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{x^2}{2h^2}\right)$$

is the normal density rescaled by the bandwidth $(N(0, h^2))$. If $h \to 0$ and $nh \to \infty$, then the optimal bandwidth is

$$h_{AMISE} = \left[\frac{1}{2\sqrt{\pi} n \int (f'')^2}\right]^{1/5}$$

where $\frac{\partial^2 f}{\partial x} = f''$. Since the optimal value is unknown, approximations methods are needed. For a derivation and discussion of these methods, see Silverman (1986) and Jones, Marron, and Sheather (1996).

General Univariate Kernel Density Estimate

Assume that $W_i = 1$, in which case the kernel density estimate for $f(x)$ is,

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h})$$

where the kernel, K, satisfies $\int K(x)dx = 1$ and the smoothing parameter, h, is called the bandwidth. In practice, the kernel is an unimodal function satisfying, $\int xK(x)dx = 0$.[7] A popular choice for the *normal* kernel given by

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

If one assumes that the underlying density is sufficiently smooth and that the kernel has finite forth moments, then an asymptotic expansion for the bias and variance of a kernel estimate are given by

$$Bias_{asy}\{\hat{f}_h(x)\} = \frac{h^2}{2}\mu_2(K)^2 f''(x)$$

and

$$Var_{asy}\{\hat{f}_h(x)\} = \frac{1}{nh} R(K) f(x)$$

where $R(K) = \int K^2(y)dy$, $\mu_2(K) = \int y^2 K(y)dy$ and f'' is the second derivative of f.

A widely used criteria for measuring the discrepancy between f and $\hat{f}$ is the mean integrated squared error (MISE) is given by,

$$\begin{aligned} MISE(\hat{f}) &= E\{\int \left(f(y) - \hat{f}(y)\right)^2 dy\} \\ &= \int Bias(\hat{f}(y))^2 dy + \int Var(\hat{f}(y)) dy. \end{aligned}$$

[7] Note: the kernel, $K(x)$ is a unimodal density function with expected value, $E_K(X) = \int xK(x)dx = 0$.

If one assumes that the function f is integrable, then the asymptotic mean integrated squared error (AMISE) is given by

$$AMISE(\hat{f}_{(h)} = \frac{1}{nh}R(K) = \frac{h^2}{4}\mu_2(K)^2 R(f'').$$

The bandwidth that minimizes the AMISE is given by

$$h_{AMISE} = \{\frac{R(K)}{\mu_2(K)^2 R(f'')}\}^{1/3} n^{-1/3}.$$

Bandwidth Selection

Several different bandwidth selection methods are available in PROC KDE in the univariate case. Following the recommendations of Jones, Marron, and Sheather (1996), the default method follows a plug-in formula of Sheather and Jones.

This method solves the fixed-point equation

$$h = \left[\frac{R(\varphi)}{nR\left(\hat{f}''_{g(h)}\right)\left(\int x^2\varphi(x)dx\right)^2}\right]^{1/5}$$

where $R(\varphi) = \int \varphi^2(x)dx$ and $g(h) = C(K)[R(f'')/R(f''')]^{1/7}h^{5/7}$ is the bandwidth for the estimate of $R(\hat{f}'')$. PROC KDE solves this equation by first evaluating it on a grid of values spaced equally on a log scale. The largest two values from this grid that bound a solution are then used as starting values for a bisection algorithm. The simple normal reference rule works by assuming $\hat{f}$ is Gaussian in the preceding fixed-point equation. This results in

$$\begin{aligned} h &= \hat{\sigma}[4/(3n)]^{1/5} \\ &= 1.06\ \hat{\sigma}n^{-1/5} \end{aligned}$$

where $\hat{\sigma}$ is the sample standard deviation.

Alternatively, the bandwidth can be computed using the interquartile range,

$$\begin{aligned} h &= 1.06\hat{\sigma}n^{-1/5} \\ &\approx 1.06\hat{\sigma}n^{-1/5} \\ &\approx 1.06\ (Q/1.34)n^{-1/5} \end{aligned}$$

Silverman's rule of thumb (Silverman, 1986, Section 3.4.2) is computed as

$$h = 0.9\min[\hat{\sigma}, Q/1.34]n^{-1/5}$$

The oversmoothed bandwidth is computed as

$$h = 3\hat{\sigma}[1/(70\sqrt{\pi}n)]^{1/5}$$

When you specify a WEIGHT variable, PROC KDE uses weighted versions of Q_3, Q_1, and $\hat{\sigma}$ in the preceding expressions. The weighted quartiles are computed as weighted order statistics, and the weighted variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n W_i (X_i - \bar{X})^2}{\sum_{i=1}^n W_i}$$

where $\bar{X} = (\sum_{i=1}^n W_i X_i)/(\sum_{i=1}^n W_i)$ is the weighted sample mean.

Bivariate Kernel Density Estimates

For the bivariate case, let $X = (X, Y)$ be a bivariate random element taking values in R^2 with joint density function $f(x, y),\ (x, y) \in R^2$ and let $X_i = (X_i, Y_i),\ i = 1, 2, \ldots, n$, be a sample of size n drawn from this distribution. The kernel density estimate of $f(x, y)$ based on this sample is

$$\begin{aligned} \hat{f}(x,y) &= \frac{1}{n}\sum_{i=1}^n \varphi_h(x - X_i, y - Y_i) \\ &= \frac{1}{nh_X h_Y}\sum_{i=1}^n \varphi\left(\frac{x - X_i}{h_X}, \frac{y - Y_i}{h_Y}\right) \end{aligned}$$

where $(x, y) \in R^2$, $h_X > 0$ and $h_Y > 0$ are the bandwidths, and $\varphi_h(x, y)$ is the rescaled normal density where $\varphi(x, y)$ is the standard normal density function

$$\varphi(x,y) = \frac{1}{2\pi}\exp\left(-\frac{x^2 + y^2}{2}\right)$$

Under mild regularity assumptions about $f(x, y)$, the mean integrated squared error (MISE) of $\hat{f}(x, y)$ is

$$\begin{aligned} \text{MISE}(h_X, h_Y) &= \text{E}\int (\hat{f} - f)^2 \\ &= \frac{1}{4\pi n h_X h_Y} + \frac{h_X^4}{4}\int \left(\frac{\partial^2 f}{\partial X^2}\right)^2 dxdy \\ &\quad + \frac{h_Y^4}{4}\int \left(\frac{\partial^2 f}{\partial Y^2}\right)^2 dxdy + O\left(h_X^4 + h_Y^4 + \frac{1}{nh_X h_Y}\right) \end{aligned}$$

as $h_X \to 0$, $h_Y \to 0$ and $nh_X h_Y \to \infty$.
Now set

$$\begin{aligned} \text{AMISE}(h_X, h_Y) &= \frac{1}{4\pi n h_X h_Y} + \frac{h_X^4}{4}\int \left(\frac{\partial^2 f}{\partial X^2}\right)^2 dxdy \\ &\quad + \frac{h_Y^4}{4}\int \left(\frac{\partial^2 f}{\partial Y^2}\right)^2 dxdy \end{aligned}$$

which is the asymptotic mean integrated squared error (AMISE). For fixed n, this has a minimum at $(h_{AMISE\_X}, h_{AMISE\_Y})$ defined as

$$h_{AMISE\_X} = \left[\frac{\int (\frac{\partial^2 f}{\partial X^2})^2}{4n\pi}\right]^{1/6}\left[\frac{\int (\frac{\partial^2 f}{\partial X^2})^2}{\int (\frac{\partial^2 f}{\partial Y^2})^2}\right]^{2/3}$$

and

$$h_{AMISE\_Y} = \left[\frac{\int (\frac{\partial^2 f}{\partial Y^2})^2}{4n\pi}\right]^{1/6}\left[\frac{\int (\frac{\partial^2 f}{\partial Y^2})^2}{\int (\frac{\partial^2 f}{\partial X^2})^2}\right]^{2/3}$$

These are the optimal asymptotic bandwidths in the sense that they minimize MISE. However, as in the univariate case, these expressions contain the second derivatives of the unknown density f being estimated, and so approximations are required. See Wand and Jones (1993) for further details.

Bandwidth Selection

For the bivariate case, Wand and Jones (1993) note that automatic bandwidth selection is both difficult and computationally expensive. Their study of various ways of specifying a bandwidth matrix also shows that using two bandwidths, one in each coordinate's direction, is often adequate. PROC KDE enables you to adjust the two bandwidths by specifying a multiplier for the default bandwidths recommended by Bowman and Foster (1993):

$$\begin{aligned} h &= \hat{\sigma}_X n^{-1/6} \\ &= \hat{\sigma}_Y n^{-1/6} \end{aligned}$$

Here $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are the sample standard deviations X and Y, respectively. These are the optimal bandwidths for two independent normal variables that have the same variances as X and Y. They are, therefore, conservative in the sense that they tend to oversmooth the surface.
You can specify the BWM= option to adjust the aforementioned bandwidths to provide the appropriate amount of smoothing for your application.

Examples

Sheather Bimodal Example - R

The R code used for this example is

```
install.packages("KernSmooth")
library(KernSmooth)

bimodal <- read.table("bimodal.txt", header=TRUE)
attach(bimodal)
x <- bimodal$x
n<-length(x)
xx <- c(-300:300)/100

sheather.curve = function(h, main=" ", sub = " ") {
plot( x=c(-3,3),y=c(0,0.65),type="n",xlab="x",ylab="Estimated & True Densities")
title(main=main, sub = sub)
ysum = numeric(601)
for (i in 1:n)
{points(x[i], 1/(n*h*sqrt(2*pi)),type="h")
x1 = numeric(601)+x[i]
y = (1/(h*sqrt(2*pi)))*exp(-0.5*((xx-x1)/h)^2)
ysum = y/n + ysum
lines(xx,y/n,lty=1)}
lines(xx,ysum,lty=1)
truedensity = 0.5*(3/(sqrt(2*pi)))*exp(-0.5*((xx+1)/(1/3))^2)
            + 0.5*(3/(sqrt(2*pi)))*exp(-0.5*((xx-1)/(1/3))^2)
lines(xx,truedensity,lty=2)
}

par(mfrow=c(2,2))
 sheather.curve(.2,"Sheather Bimodal Data", "with smoother = .2")
```

```
sheather.curve(.4,"", "with smoother = .4")
sheather.curve(.6,"", "with smoother = .6")
sheather.curve(.8,"", "with smoother = .8")
```

The output is given in Figure 1.5 Two SAS output plots for the bimodal data are given in Figure 1.6

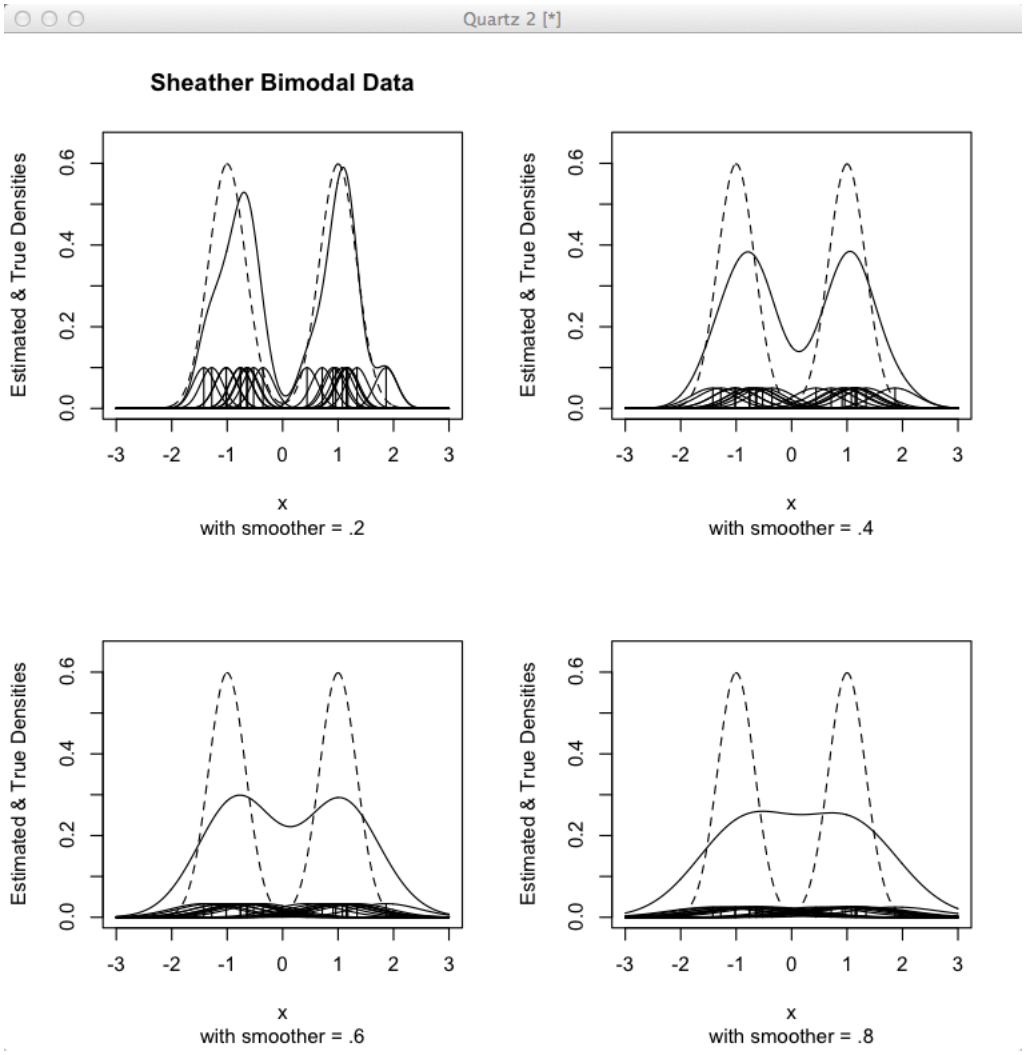


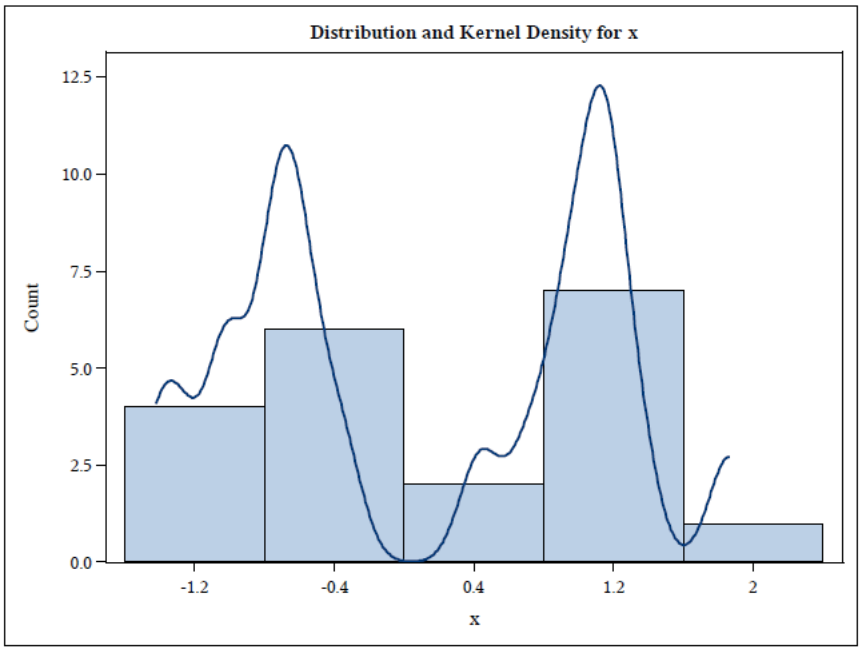
Figure 1.5: Sheather Bimodal Example

Old Faithful Example

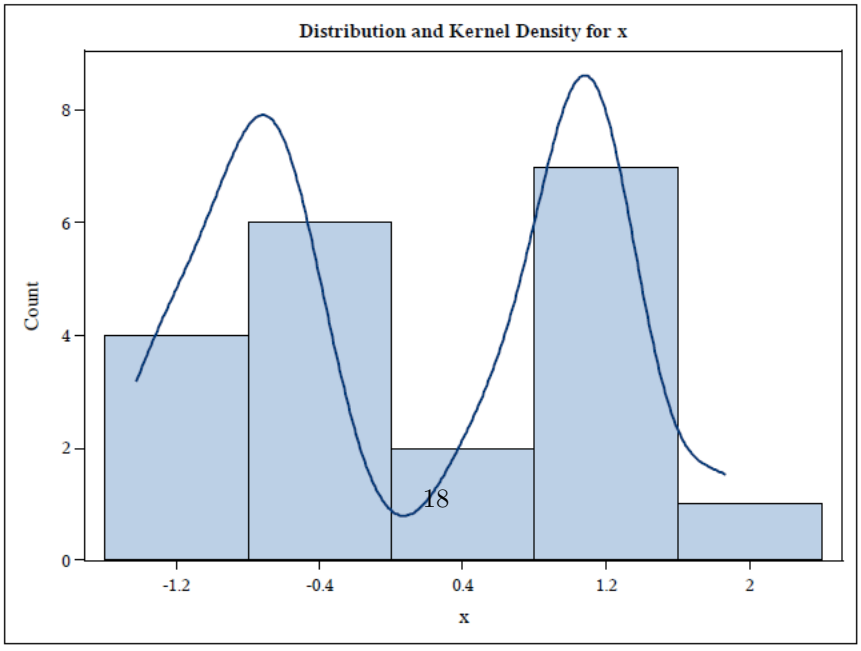
The R code is

```
# The Old Faithful geyser data
par(mfrow=c(2,1))
 library(KernSmooth)
```

Controls	
	x
Grid Points	401
Lower Grid Limit	-1.421
Upper Grid Limit	1.8661
Bandwidth Multiplier	0.4



Controls	
	x
Grid Points	401
Lower Grid Limit	-1.421
Upper Grid Limit	1.8661
Bandwidth Multiplier	0.8



```
attach(faithful)
hist(x=waiting)
fhat <- bkde(x=waiting)
plot (fhat, xlab="x", ylab="Density function")
```

The output is given in Figure 1.7

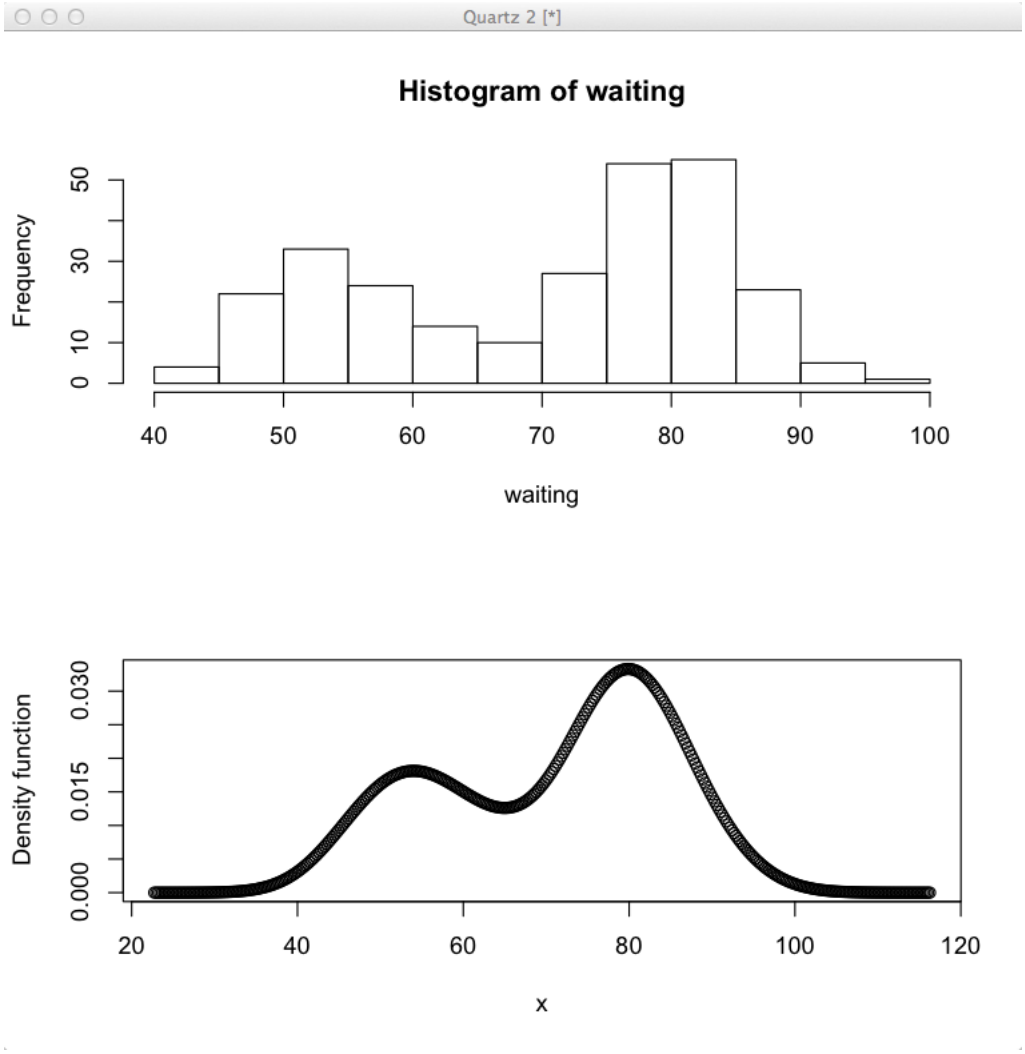


Figure 1.7: Old Faithful Waiting Times

The SAS output is given in Figure 1.8

Old Faithful Data

The KDE Procedure

Inputs	
Data Set	WORK.FAITHFUL
Number of Observations Used	272
Variable	wait
Bandwidth Method	Sheather-Jones Plug In

Controls	
	wait
Grid Points	401
Lower Grid Limit	43
Upper Grid Limit	96
Bandwidth Multiplier	1

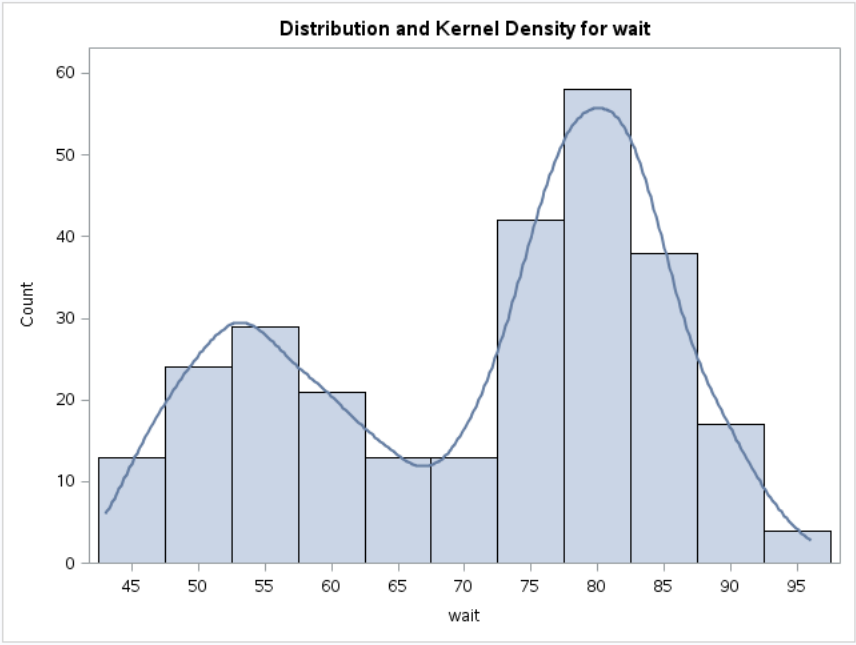


Figure 1.8: Old Faithful Waiting Times

1.3.4 Box-Cox Transformations

Suppose that $y > 0$, define the Box-Cox transformation[8] as

$$y_i^{(\lambda)} = \begin{cases} (y_i^\lambda - 1)/\lambda & when \quad \lambda \neq 0, \\ \ln y_i & when \quad \lambda = 0, \end{cases}$$

[8] This transformation tends to be overused! Do not transform the data unless you have a very good reason for doing so. Just making it look "more normal" may not be sufficient.

where $i = 1, 2, \ldots, n$. One determines λ by maximizing

$$-n/2 \log[s^2(\lambda)] = (\lambda - 1) \sum_{i=1}^{n} \ln(y_i) - n/2 \log[\hat{\sigma}^2(\lambda)],$$

and $\hat{\sigma}^2(\lambda) = 1/n \vec{y}^{(\lambda)\prime}[I - H]\vec{y}^{(\lambda)}$ i.e., it is the sum of squares for the error term when $y_i^{(\lambda)}$ is used instead of y_i and $\vec{y}^{(\lambda)} = (y_1^{(\lambda)}, y_2^{(\lambda)}, \ldots, y_n^{(\lambda)})'$.
Since, there is not a close form solution to the above maximization, one usually plots $-n/2 \log[s^2(\lambda)]$ vs λ. Another approach is to compute a confidence interval using the fact that $-n/2 \log[s^2(\lambda)] \sim \chi^2(df = 1)$. One can then use any λ which is contained in the confidence interval.

Examples

The SAS code is

```
title 'Univariate Box-Cox';
data x;
   call streaminit(17);
   z = 0;
   do i = 1 to 500;
      y = rand("LOGNORMAL");
      output;
   end;
run;
proc transreg maxiter=0 nozeroconstant;
   model BoxCox(y) = identity(z);
   output;
run;
proc univariate noprint;
   histogram y ty;
run;
```

The output is given in Figure 1.9. The R-Code is

```
install.packages("MASS")
library(MASS)
par(mfrow=c(3,1))
y = rlnorm(500)
x = rnorm(500)
boxcox(y ~ x, lambda = seq(-2, 2, length = 100))
hist(y)
hist(log(y))
```

The R output is given in Figure 1.10

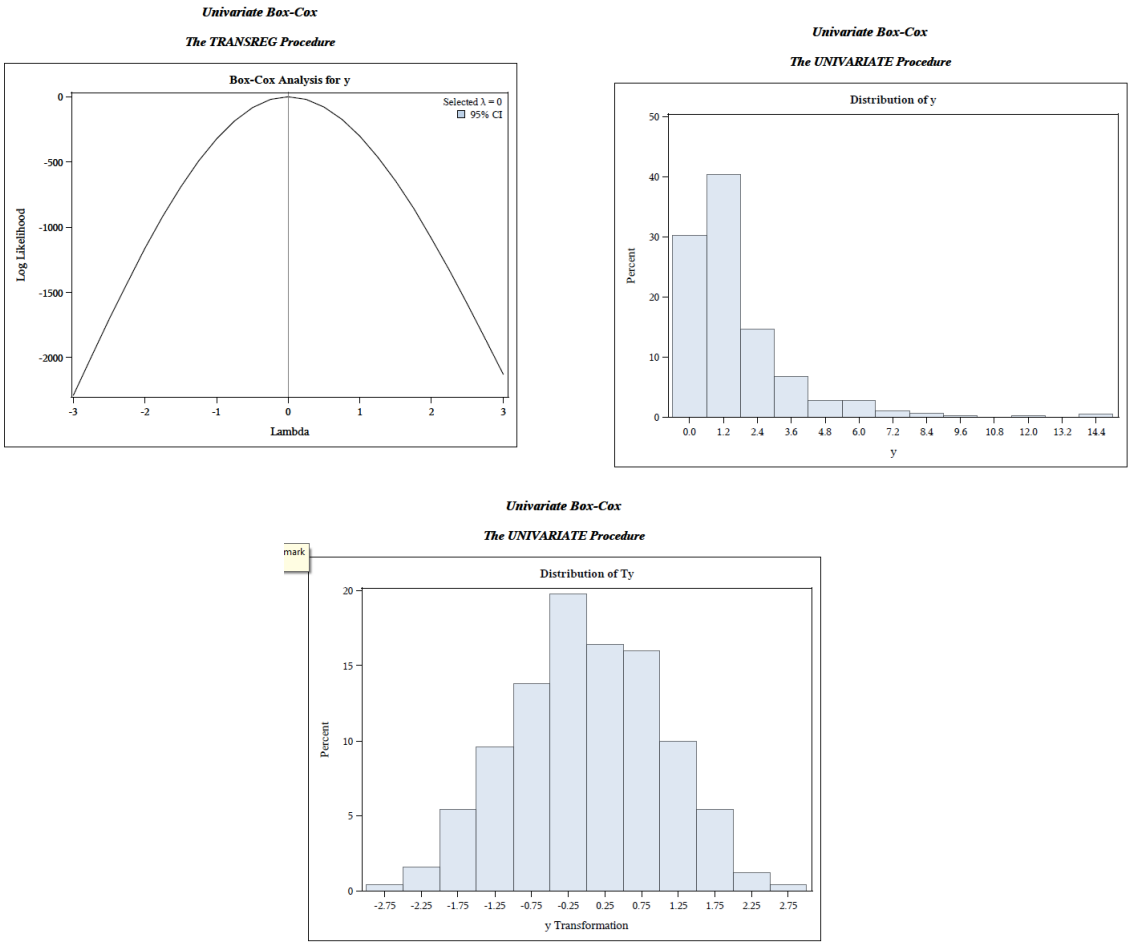


Figure 1.9: Univariate Box-Cox Transformation

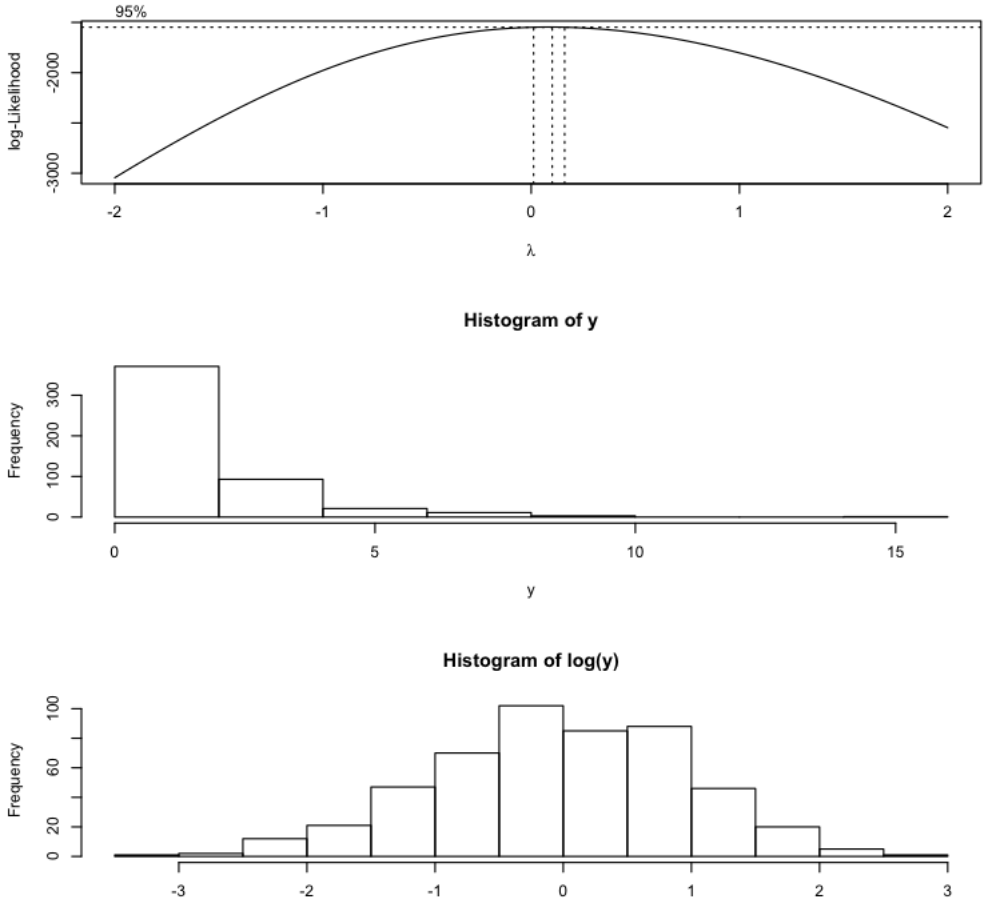


Figure 1.10: Univariate Box-Cox Transformation

1.3.5 Bootstrap Methods

This topic will be covered in much greater detail in an advanced computational statistics course. Here, I am using this numerical computational procedure to determine the standard error of a statistic. This method is particularly useful when the derivation of the standard error is mathematically complicated.
Suppose that one has a realization of a simple random sample of size n, given by $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. A single bootstrap sample given by, $\mathbf{x}_b^* = (x_1^*, x_2^*, \ldots, x_n^*)$, is a sample of size n taken from the above realization when the sampling is done **with replacement**. Suppose that T_n is any statistic. The standard error of T_n can be computed as:

1. Select B (large number) independent bootstrap samples $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_B^*$, where $\mathbf{x}_b^* = (x_1^*, x_2^*, \ldots, x_n^*)$.
2. Compute the statistic $T_n = T_n^{*b}$ for each of the $b = 1, 2, \ldots, B$ bootstrap samples.
3. Estimate the bootstrap standard error by

$$s.e._{boot}(T_n) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(T_N^{*b} - \bar{T}_n^* \right)^2 \right\}^{1/2}$$

where $\bar{T}_n^* = \frac{1}{B} \sum_{b=1}^{B} T_N^{*b}$.

Bootstrap Example

R provides an easy method of simulating the bootstrap methods. [You will need to install the R package *simple.boot*]. The R code is

```
set.seed(20)
theta = 12  # parameter for the uniform (0, theta)
x <- runif(25)
x = x*theta
b.mean <- one.boot(x, mean, 100)
b.mean$t = 2*b.mean$t
sd(b.mean$t)
boxplot(b.mean$t)

## The statistics is the quantile (1.0 is the max)
b.max <-one.boot(x, quantile, R = 100, probs = 1.0)
sd(b.max$t)
boxplot(b.max$t)
```

The output is

```
> set.seed(20)
> theta = 12
> x <- runif(25)
> x = x*theta
> b.mean <- one.boot(x, mean, 100)
> b.mean$t = 2*b.mean$t
> sd(b.mean$t)
[1] 1.358726
> boxplot(b.mean$t)
>
```

```
> ## The statistics is the quantile (1.0 is the max)
> b.max <-one.boot(x, quantile, R = 100, probs = 1.0)
> sd(b.max$t)
[1] 0.5818098
> boxplot(b.max$t)
```

In this example ($\theta = 12$) the standard error for $2\bar{x}$ is 1.358 (1.92) and the standard error for the n^{th} order statistic is 0.58 (0.43) when $n = 25$ and B = 100. The actual value is in parenthesize.

Chapter 2

Two Population Methods

2.1 Discrete Variables

2.1.1 Contingency Tables

Two Way (r $\times$ c) Contingency Tables

Suppose that one has two discrete random variables, X with r categories and Y with c categories. Let $\pi_{ij} = \Pr[X = i, Y = j] = \Pr[a_{ij} = 1]$ denote the joint probability for the ij^{th} cell where $a_{ij} = I_i(X) \times I_j(Y)$ and $I_j(Z) = 1$ if $Z = j$ and $I_j(Z) = 0$ if $Z \neq j$. The marginal probabilities for X and Y are,

$$\pi_{i+} = \sum_j \pi_{ij} = \Pr[X = i]$$

and

$$\pi_{+j} = \sum_i \pi_{ij} = \Pr[Y = j].$$

Suppose that one has n observations that are categorized jointly by X and Y. Let $n_{ij} = \sum a_{ij}$, $n_{i+} = \sum_j n_{ij}$, $n_{+j} = \sum_i n_{ij}$, in which case, one has $\sum_i n_{i+} = \sum_j n_{+j} = \sum_i \sum_j n_{ij} = n$. The possible outcomes form a $r \times c$ contingency or cross-classification table. The entries can be summarized in the table

Group	1	2	$\cdots$	c	Total
1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+c}	n

Independence of Categorical Variables

Definition: 2.1 Independence *The two categorical random variables X and Y are said to be independent if and only if, $\pi_{ij} = \pi_{i+}\ \pi_{+j}$ for all values i and j.*

The statistics given in the next section can be used to test the assumption that X and Y are independent. The null distribution for these statistics is often a chi-square distribution.

Chi-square Statistics

Pearson's Chi-square Statistic

The familiar Pearson chi-square statistic is given by,

$$Q_p = \sum_i \sum_j (n_{ij} - m_{ij})^2 / m_{ij}.$$

Where Q_p has an asymptotic null chi-square distribution with degrees of freedom $= (r-1) \times (c-1)$ and $m_{ij} = E(n_{ij} \mid H_0) = \frac{n_{i+}\ n_{+j}}{n}$.

Likelihood-Ratio Test

The likelihood-ratio statistic involves the ratios between the observed and expected frequencies. The statistic is computed as

$$G^2 = Q_L = 2 \sum_i \sum_j n_{ij}\ log\ [n_{ij}/m_{ij}].$$

This statistic is seldom used in the 2×2 tables.

Mantel-Haenszel Test

The Mantel-Haenszel statistic is used when testing the null hypothesis versus the alternative hypothesis that there is a linear association between the row variable and the column variable. **Both variables must be at least ordinal variables.** The statistic is computed as,

$$Q_{MH} = (n-1)\hat{\rho}$$

where $\hat{\rho}$ is the Pearson correlation between the row variable and the column variable. This statistics is more widely used when there are multiple tables.

Measures of Association

Phi Coefficient

The phi coefficient is a measure of association derived from the Pearson chi-square statistic. It has the range $\phi \in [-1, 1]$ for 2×2 tables. Otherwise, the range is $0 \leq \phi \leq min[\sqrt{(r-1)}, \sqrt{(c-1)}]$, (Liebetrau 1983). The phi coefficient is computed as

$$\phi = \frac{n_{11}\ n_{22} - n_{12}\ n_{21}}{n_{1+}\ n_{2+}\ n_{+1}\ n_{+2}}$$

when $r = c = 2$ and is,

$$\phi = \sqrt{Q_p / n}$$

when either r or c are not equal to 2.

Contingency Coefficient

The contingency coefficient is a measure of association derived from the Pearson chi-square. It has the range $0 \leq P \leq \sqrt{(m-1)/m}$, where $m = min(r, c)$ (Liebetrau 1983). The contingency coefficient is computed as,

$$P = \sqrt{\frac{Q_p}{Q_p + n}}.$$

Refer to Kendall and Stuart (1979, pp. 587 -588).

Cramer's V

Cramer's V is a measure of association derived from the Pearson chi-square. It is designed so that the attainable upper bound is always 1. It has the range $V \in [-1, 1]$ for 2×2 tables; otherwise, the range is $V \in [0, 1]$. Cramer's V is computed as

$$V = \phi,$$

when $r = c = 2$ and otherwise

$$V = \sqrt{Q_p / nm},$$

where $m = min(r - 1, c - 1)$. Refer to Kendall and Stuart (1979, p. 588).

Exact Tests

Whenever the sample size is small it is appropriate to consider the exact test. Assume that the marginal frequencies are fixed, in which case the data (row or column cell frequencies) are distributed as a hypergeometric distribution. The probability of the observed cell frequency are

$$\Pr[n_{ij}] = \frac{n_{1+}!\ n_{2+}!\ n_{+1}!\ n_{+2}!}{n!\ n_{11}!\ n_{12}!\ n_{21}!\ n_{22}!},$$

or

$$p(t) = \Pr[n_{11} = t] = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}}.$$

This method is best illustrated by using an example.

Suppose one observes the following table,

Group	f	u	Total
test	10	2	12
placebo	2	4	6
Total	12	6	18

The possible tables with these marginal totals are

(1,1)	(1,2)	(2,1)	(2,2)	Probabilities
12	0	0	6	0.0001
11	1	1	5	0.0039
10	2	2	4	0.0533
9	3	3	3	0.2370
8	4	4	2	0.4000
7	5	5	1	0.2560
6	6	6	0	0.0498

To find the one sided p-value sum the probabilities as small or smaller than those computed for the observed table, in the direction specified by the one-sided alternative. In this case, it would be those values where test would be more favorable as (sum of the red probabilities)

$$p = 0.0533 + 0.0039 + 0.0001 = 0.0573.$$

To find the two sided p-value sum all the probabilities as small or smaller than the observed value (sum of the red and blue probabilities)

$$p = 0.0533 + 0.0039 + 0.0001 + 0.0498 = 0.1071.$$

In this next section, consider the special case when $r = 2$ and $c = 2$.

Special Case: the (2×2) Tables

In this section, we consider a special case of the general $r \times c$ tables where both X and Y are binary random variables ($r = 2$ and $c = 2$). In addition, we will consider the circumstance when one of the categorical variables, say X, is not random (e.g., X specifies gender). In this situation, the conditional probability of category $Y = j$ given $X = i$, $\pi_{j|i}$, is the parameter of interest. Consider the 2×2 table where one is interested in comparing $\Pr[Y = 1 \mid X = 1] = \pi_1 = \pi_{1|1}$ and $\Pr[Y = 1 \mid X = 2] = \pi_2 = \pi_{1|2}$ when $Y = 1$ is an event of interest The response variable Y is statistically independent of the row classification, X, when $\pi_1 - \pi_2 = 0$. This concept of computing this difference works well when $r = 2$ but doesn't when $r > 2$. In which case, the following ratios are commonly used.

Relative Risk

The relative risk is

$$RR = \frac{\Pr[[Y = 1 \mid X = 1]}{\Pr[Y = 1 \mid X = 2]} = \frac{\pi_1}{\pi_2}. \tag{2.1}$$

Odds Ratio

The odds of an success is

$$\text{Odds of success} = \frac{\pi}{(1 - \pi)}. \tag{2.2}$$

The odds ratio of an success for the two rows defined by X is

$$OR = \theta = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)} \tag{2.3}$$

when X is not random. The odds ratio is

$$OR = \theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}. \tag{2.4}$$

when X and Y are random, Note: In the conditional case, independence implies that RR and $OR = 1$.

Inference for 2×2 Tables

Suppose one has the following table where the row variable X is the random assignment of subject to either the control (placebo) or treatment (active) groups and Y denotes whether or not there is a "favorable (f)" or "unfavorable (u)" outcome.

	f	u	Total
active	n_{11}	n_{12}	n_{1+}
placebo	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

If one was interested in testing the null hypothesis that there is no association between the treatment and the outcome of the treatment and the marginal totals are fixed then it follows that,

$$\Pr[n_{ij}] = \frac{n_{1+}!\ n_{2+}!\ n_{+1}!\ n_{+2}!}{n!\ n_{11}!\ n_{12}!\ n_{21}!\ n_{22}!},$$

and

$$E(n_{ij} \mid H_0) = \frac{n_{i+}\ n_{+j}}{n} = m_{ij}, \quad V(n_{ij} \mid H_0) = \frac{n_{1+}\ n_{2+}\ n_{+1}\ n_{+2}}{n^2(n-1)} = v_{ij}.$$

When the total sample size n is sufficiently large, n_{11} is a sufficient statistic and has an approximate normal distribution from which one has,

$$Q = \frac{(n_{11} - m_{11})^2}{v_{11}} \sim \chi^2(df = 1),$$

and

$$Q_p = \sum_{i=1}^{2}\sum_{j=1}^{2}(n_{ij} - m_{ij})^2/m_{ij} = \frac{n}{n-1}\ Q.$$

It can be shown that the Pearson correlation coefficient, $\hat{\rho}$ is related to Q_p by

$$\hat{\rho} = [n_{1+}\ n_{2+}/n_{+1}\ n_{+2}]^{1/2}\ (\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{Q_p/n}.$$

Inference for Difference in Proportions

Suppose one wants to test the hypothesis that the probability of a favorable outcome given the active treatment, Pr[f | active] $= \pi_{1|1} = \pi_{11}/\pi_{1+} = \pi_1$, is the same as the probability of having a favorable outcome using the placebo, Pr[f | placebo] $= \pi_{1|2} = \pi_{21}/\pi_{2+} = \pi_2$. This hypothesis is denoted as $H_0 : \pi_1 = \pi_2$. Define $\hat{\pi}_1 = n_{11}/n_{1+}$ and $\hat{\pi}_2 = n_{21}/n_{2+}$ in which case it follows that $E[\hat{\pi}_1 - \hat{\pi}_2] = \pi_1 - \pi_2$ and $Var[\hat{\pi}_1 - \hat{\pi}_2] = \pi_1(1-\pi_1)/n_{1+} + \pi_2(1-\pi_2)/n_{2+}$. Using an unbiased estimate of $Var[\hat{\pi}_1 - \hat{\pi}_2]$ given by

$$v_d = \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{(n_{1+}-1)} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{(n_{2+}-1)}$$

allows one to define a 100(1-α)% confidence interval for $(\pi_1 - \pi_2)$ as

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm \{z_{\alpha/2}\ \sqrt{v_d}\}$$

or

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm \{z_{\alpha/2}\ \sqrt{v_d} + [1/2(1/n_{1+} + 1/n_{2+})]\}.$$

Inference for Odds Ratio and Relative Risk

Relative Risk

From equation (2.1) the relative risk is given as $RR = \pi_1/\pi_2$ where $\Pr[Y = 1 \mid X = 1] = \pi_1 = \pi_{1|1}$ and $\Pr[Y = 1 \mid X = 2] = \pi_2 = \pi_{1|2}$. An estimate for the relative risk is, $\widehat{rr} = \hat{\pi}_1/\hat{\pi}_2$. The asymptotic properties for the log of the relative risk are easier to derive than for the relative risk, in which case, the estimated standard error for the relative risk, $\log(rr)$, is

$$\hat{\sigma}_{log(rr)} = [\frac{(1-\hat{\pi}_1)}{\hat{\pi}_1\ n_{1+}} + \frac{(1-\hat{\pi}_2)}{\hat{\pi}_2\ n_{2+}}]^{1/2}.$$

The Wald confidence interval is

$$log(rr) \pm z_{\alpha/2}\ \hat{\sigma}_{log(rr)}.$$

Odds Ratio

From equation (2.4) the odds ratio is $OR = \theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$. An estimate for the odds ratio is

$$\widehat{OR} = \hat{\theta} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Note, since $\theta = \infty$ if either n_{12} or n_{21} equal zero [this can happen with positive probability]. An alternative estimate for the odds ratio is given by

$$\tilde{\theta} = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)}.$$

$\hat{\theta}$ and $\tilde{\theta}$ have the same asymptotic distribution but neither are well behaved for small n. As in the relative risk, the log of the odds ratio has better asymptotic properties. The estimated standard error for the log odds ratio is

$$\hat{\sigma}_{log(\hat{\theta})} = [\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}]^{1/2}.$$

The Wald confidence interval for $log(\theta)$ is

$$log(\hat{\theta}) \pm z_{\alpha/2}\ \hat{\sigma}_{log(\hat{\theta})}.$$

Note: the computation and asymptotic normality of the log odds and log relative risk follow from the Delta method [1]

Delta Method for the Log Odds Ratio

The asymptotic standard errors for the log relative risk and the log odds ratio are derived using the same underlying multinomial distribution with the multi-parameter version of the delta method.

Suppose that $\{n_i, i = 1, 2, \ldots, c\}$ have a multinomial $(n, \{\pi_i\})$ distribution. Let $\hat{\pi}_i = n_i/n$ where one has

$$E(\hat{\pi}_i) = \pi_i \qquad Var(\hat{\pi}_i) = \pi_i(1-\pi_i)/n \qquad cov(\hat{\pi}_i, \hat{\pi}_j) = -\pi_i\pi_j/n.$$

The sample estimates $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{c-1})$ have an asymptotic normal distribution. Let $g(\pi)$ denote a differential function of $\pi = (\pi_1, \pi_2, \ldots, \pi_{c-1})$ with samples $g(\hat{\pi})$. Let

$$\phi_i = \frac{\partial g(\pi)}{\partial \pi_i}$$

for $i = 1, 2, \ldots, c-1$. Then

$$\sqrt{n}[g(\hat{\pi}) - g(\pi)]/\sigma \to N(0,1) \tag{2.5}$$

as $n \to \infty$, where

$$\sigma^2 = \sum_i \pi_i\phi_i^2 - (\sum_i \pi_i\phi_i)^2. \tag{2.6}$$

When $g(\pi)$ is the log odds ratio, we have

$$g(\pi) = \log \pi_{11} + \log \pi_{22} - \log \pi_{12} - \log \pi_{21}$$

and

$$\phi_{ii} = 1/\pi_{ii}$$

[1] Agresti (edition 2) pages 73-77.

and

$$\phi_{ij} = -1/\pi_{ij}$$

for $i, j = 1, 2$. In which case

$$\begin{aligned} \hat{\sigma}^2 &= \sum_i \pi_i \phi_i^2 - (\sum_i \pi_i \phi_i)^2 \\ &= \sum_i \sum_j 1/n\hat{\pi}_{ij} \\ &= \sum_i \sum_j 1/n_{ij} \end{aligned} \tag{2.7}$$

where $\sum_i \pi_i \phi_i$ in equation(2.6) is

$$\sum_i \sum_j \pi_{ij} \phi_{ij} = 0,$$

and $\sum_i \pi_i \phi_i^2$ in equation(2.6) is

$$\sum_i \sum_j \pi_{ij} \phi_{ij}^2 = \sum_i \sum_j 1/\pi_{ij},$$

and $n\hat{\pi}_{ij} = n_{ij}$.

Mantel-Haenszel Test

Suppose that one has q independent 2×2 tables where one assumes that the marginal sums are fixed. These assumptions insure that one has a hypergeometric distribution. It follows that,

$$E(n_{hij} \mid H_0) = \frac{n_{hi+}\ n_{h+j}}{n_h} = m_{hij} \text{ and } V(n_{hij} \mid H_0) = \frac{n_{h1+}\ n_{h2+}\ n_{h+1}\ n_{h+2}}{n_h^2(n_h - 1)} = v_{hij}.$$

From which the Mantel-Haenszel test statistic is given by

$$Q_{MH} = \frac{[\sum_{h=1}^q n_{h11} - \sum_{h=1}^q m_{h11}]^2}{\sum_{h=1}^q v_{h11}}.$$

Homogeneity of Odds Ratios

The Breslow-Day statistic is given by

$$Q_{BD} = \sum_h^q \sum_i^2 \sum_j^2 \frac{(n_{hij} - m_{hij})^2}{m_{hij}}, \tag{2.8}$$

which has a asymptotic chi-square distribution with $q - 1$ degrees of freedom.

The SAS USER's guide has the following concerning the Breslow-Day procedure.

Breslow-Day Test for Homogeneity of the Odds Ratios

When you specify the CMH option, PROC FREQ computes the Breslow-Day test for stratified analysis of 2×2 tables. It tests the null hypothesis that the odds ratios for the q strata are all equal. When the null hypothesis is true, the statistic has approximately a chi-square distribution with $q - 1$ degrees of freedom. Refer to Breslow and Day (1980) and Agresti (1996).

The Breslow-Day statistic is computed as

$$Q_{BD} = \sum_h \frac{(n_{h11} - E(n_{h11}|OR_{MH}))^2}{var(n_{h11}|OR_{MH})}.$$

For the Breslow-Day test to be valid, the sample size should be relatively large in each stratum, and at least 80% of the expected cell counts should be greater than 5. Note that this is a stricter sample size requirement than the requirement for the Cochran-Mantel-Haenszel test for $q \times 2 \times 2$ tables, in that each stratum sample size (not just the overall sample size) must be relatively large. Even when the Breslow-Day test is valid, it may not be very powerful against certain alternatives, as discussed in Breslow and Day (1980).

If you specify the BDT option, PROC FREQ computes the Breslow-Day test with Tarone's adjustment, which subtracts an adjustment factor from Q_{BD} to make the resulting statistic asymptotically chi-square.

$$Q_{BDT} = Q_{BD} - \frac{(\sum_h (n_{h11} - E(n_{h11}|OR_{MH})))^2}{\sum_h var(n_{h11}|OR_{MH})}$$

Refer to Tarone (1985), Jones et al. (1989), and Breslow (1996).

2.1.2 Binary or Diagnostic Tests

One of the applications of 2×2 contingency tables is found in the diagnostic testing literature[2]. The material given in this section has been taken from M. S. Pepe's text, "The Statistical Evaluation of Medical Tests for Classification and Prediction". Suppose that a diagnostic test Y is binary where $Y = 1$ if the test indicates a disease and $Y = 0$ if the test indicates the absence of a disease. Let the random variable D indicate the true disease state, that is, $D = 1$ if the subject has the disease and $D = 0$ if the subject does not have the disease. The possible results are given in the classification table,

	D = 0	D = 1
Y = 0	True negative (TN)	False negative (FN)
Y = 1	False positive (FP)	True positive (TP)

A test can produce errors of two types:

$$\text{False positive fraction} = \text{FPF} = \Pr[Y = 1 \mid D = 0] \tag{2.9}$$
$$\text{False negative fraction} = \text{FNF} = \Pr[Y = 0 \mid D = 1]. \tag{2.10}$$

Additional notation is often given as

$$\begin{aligned} \text{test sensitivity} &= \text{TPF} = \Pr[Y = 1 \mid D = 1] \\ \text{test specificity} &= 1 - \text{FPF} = \Pr[Y = 0 \mid D = 0] \\ \text{disease prevalence} &= \rho = \Pr[D = 1]. \end{aligned}$$

Note: an ideal test would have TPF = 1 and FPF = 0 whereas a worthless test would have TPF = FPF that is, $\Pr[Y = 1 \mid D = 1] = \Pr[Y = 1 \mid D = 0]$. For this reason one can plot the pair (FPF, TPF) on the usual (x, y) axis. Since these values are probabilities, the pair is constrained to lie in the box with vertexes

[2]Although the material found in this section is commonly used when creating or evaluating screening or diagnostic tests, such as pap smears, PSA levels, HIV, mammograms. It has been a topic of great interest since the early months in 2020 with the onset of Covid-19 and the presence of SARS coV-2 virus or the presence of anti-bodies to the infection caused by this pathogen. In fact, we have all been forced to learn and practice critical steps in the control or mitigation of infectious diseases and pandemic outbreaks.

$(0,0),(0,1),(1,0),(1,1)$ with the ideal test lying on the point $(0,1)$ and the point for the worthless test lying on the diagonal line connecting $(0,0)$ with $(1,1)$.

The probability of misclassification, given by

$$\Pr[Y \neq D] = \rho \text{ (1 - TPF)} + (1-\rho) \text{ (FPF)}$$

is highly dependent upon the disease prevalence ρ.

Predictive Values

A commonly used probability for evaluating a test is its predictive probability of a correct decision, given by

$$\text{positive predictive value} = \text{PPV} = \Pr[D = 1 \mid Y = 1] \tag{2.11}$$
$$\text{negative predictive value} = \text{NPV} = \Pr[D = 0 \mid Y = 0]. \tag{2.12}$$

A perfect test would have PPV = 1 and NPV = 1, whereas a worthless test would not provide any additional information over what is already known in the population. That is,

$$\text{PPV} = \Pr[D = 1 \mid Y = 1] = \Pr[D = 1] = \rho$$

and

$$\text{NPV} = \Pr[D = 0 \mid Y = 0] = \Pr[D = 0] = (1-\rho).$$

One can derive the following using Bayes formula when the probability of a positive test is given by $\tau = \Pr[Y = 1]$:

$$\begin{aligned} \tau &= \rho \text{ TPF} + (1-\rho) \text{ FPF} \\ \text{PPV} &= \rho \text{ TPF}/[\rho \text{ TPF} + (1-\rho) \text{ FPF}] = \rho \text{ TPF}/\tau \\ \text{NPV} &= (1-\rho) \text{ (1 - FPF)}/[(1-\rho) \text{ (1 - FPF)} + \rho \text{ (1 - TPF)}] \end{aligned}$$

and

$$\begin{aligned} \text{TPF} &= \tau \text{ PPV}/[\tau \text{ PPV} + (1-\tau) \text{ (1 - NPV)}] \\ \text{FPF} &= \tau \text{ (1 - PPV)}/[(\tau \text{ (1 - PPV)} + (1-\tau) \text{ NPV}] \\ \rho &= \tau \text{ PPV} + (1-\tau) \text{ (1 - NPV)}. \end{aligned}$$

Example

Consider the example where the probabilities are assumed to be known.

	D = 0	D = 1	
Y = 0	.223	.142	.365
Y = 1	.078	.556	.634
	.301	.698	1.00

From which one has

$$\begin{aligned} \text{TPF} &= 0.797, \quad \text{FPF} = 0.259, \quad \rho = 0.698 \\ \text{PPV} &= 0.877, \quad \text{NPV} = 0.611, \quad \tau = 0.634. \end{aligned}$$

In the next section, a graphical method for summarizing the above probabilities is given. The curve is called the Receiver Operating Curve (ROC).

2.1.3 Receiver Operating Characteristics Curve (ROC)

In this section, assume that the random variable Y is continuous and that the test is said to be positive if $Y \geq c$, for some c. For example, let Y denote the PSA levels that is commonly used to indicate potential problems with the prostrate gland when Y is "large". The binary test given in the previous section can be constructed for any value of c. That is, the test is positive if $Y \geq c$ and is negative if $Y < c$, from which we have

$$\begin{aligned} \text{FPF}(c) &= \Pr[Y \geq c \mid D = 0] \\ \text{TPF}(c) &= \Pr[Y \geq c \mid D = 1]. \end{aligned}$$

The receiver operating characteristic curve (ROC) for a test using the random variable Y is defined as

$$ROC(\cdot) = \{(\text{FPF}(c), \text{TPF}(c)), c \in (-\infty, \infty)\} \tag{2.13}$$

or

$$ROC(\cdot) = \{(t, ROC(t)), t \in (0, 1)\}. \tag{2.14}$$

Some of the properties for the ROC include:

1. The ROC curve is invariant to strictly (monotone) increasing transformations of Y
2. Let $S_D = 1 - F_Y(y \mid D = 1)$ and $S_{\bar{D}} = 1 - F_Y(y \mid D = 0)$ denote the survivor functions of Y for the diseased and non-diseased populations given by

   $$S_D(y) = \Pr[Y \geq y \mid D = 1]$$

   $$S_{\bar{D}}(y) = \Pr[Y \geq y \mid D = 0]$$

   then

   $$ROC(t) = S_D(S_{\bar{D}}^{-1}(t)), t \in (0, 1).$$

3. $$\frac{\partial ROC(t)}{\partial t} = \frac{f_D(S_{\bar{D}}^{-1}(t))}{f_{\bar{D}}(S_D^{-1}(t))}$$

   where f_D denotes the probability density for Y in the diseased population ($D = 1$) and $f_{\bar{D}}$ denotes the probability density for Y in the healthy population ($D = 0$).

4. The area under the ROC curve (AUC) is

   $$\begin{aligned} AUC &= \Pr[Y_D > Y_{\bar{D}}] && (2.15) \\ &= \int ROC(t)dt. && (2.16) \end{aligned}$$

5. (Special Case - Parametric Binormal Form) Suppose that $Y_D \sim N(\mu_D, \sigma_D^2)$ and $Y_{\bar{D}} \sim N(\mu_{\bar{D}}, \sigma_{\bar{D}}^2)$ then

   $$ROC(t) = \Phi(a + b\Phi^{-1}(t))$$

   and

   $$AUC = \Phi(\frac{a}{\sqrt{1 + b^2}})$$

   where $a = \frac{\mu_D - \mu_{\bar{D}}}{\sigma_D}$, $b = \frac{\sigma_{\bar{D}}}{\sigma_D}$ and Φ is the standard normal c.d.f.

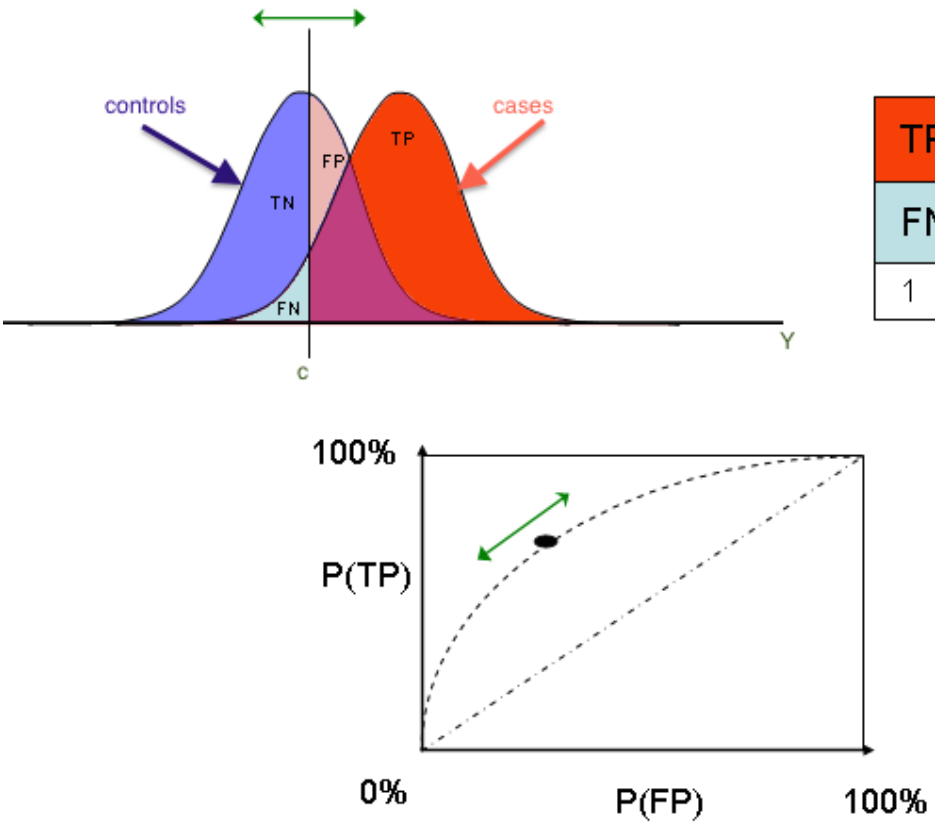


Figure 2.1: ROC Basics

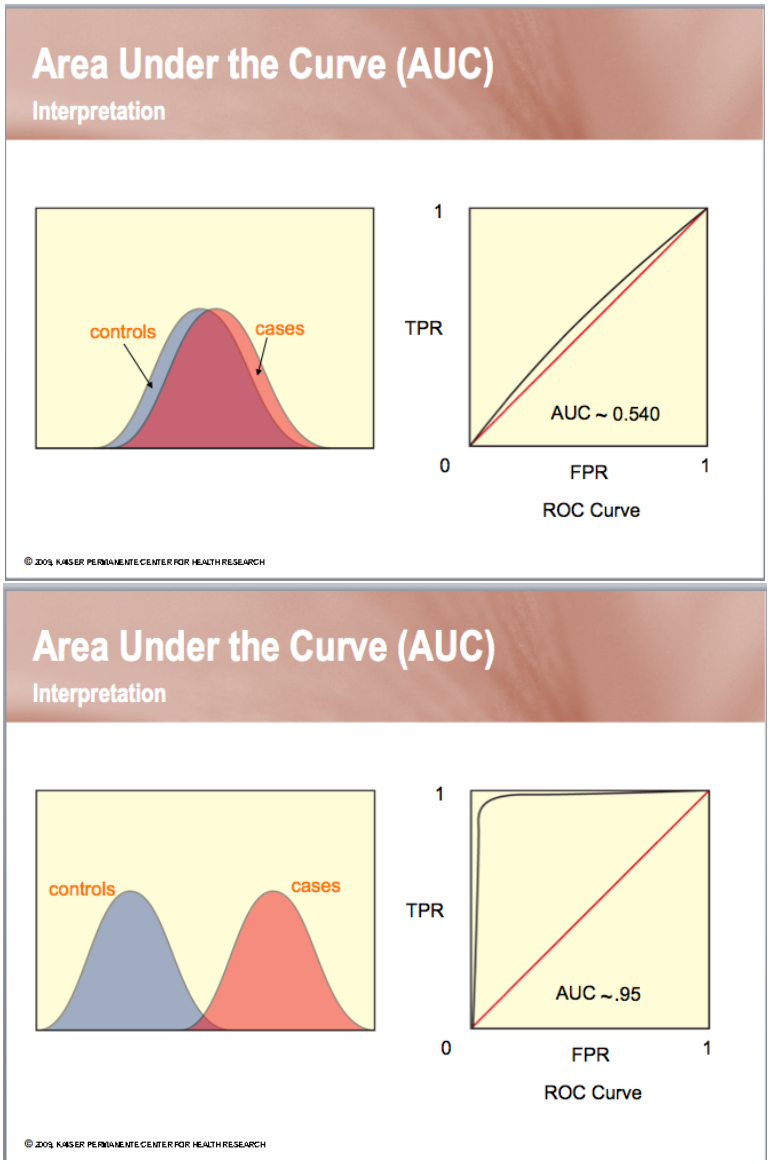


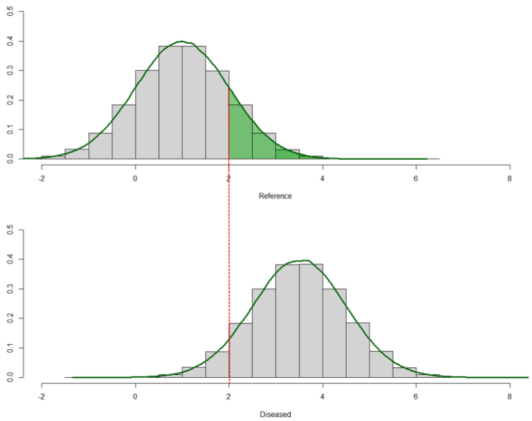
Figure 2.2: Area Under the Curve - AUC

Placement Values

A useful concept that is related to the ROC and AUC is a value called the Placement score.

Placement Values

- Cai(2002) defines $PV_D = S_{\bar{D}}(Y_D)$.



- The ROC is equivalent to the cdf of PV_D.

$$\begin{aligned} P[PV_D \leq t|\boldsymbol{X}] &= P[S_{\bar{D}\boldsymbol{X}}(Y_D) \leq t|\boldsymbol{X}] \\ &= P[Y_D \geq [S^{-1}_{\bar{D}\boldsymbol{X}}(t)|\boldsymbol{X}] \\ &= ROC_{\boldsymbol{X}}(t). \end{aligned}$$

Figure 2.3:

Stata – Example

In this example two variables; fpsa and tpsa are used as possible indicators of prostrate cancer. This data can be found at (Pancreatic Ca biomarkers)

`http://labs.fhcrc.org/pepe/book/#datasets`

Which variable is the better predictor? The following example can be summarized as Table 2.1

Table 2.1: Summary statistics

Variable	Mean	Std. Dev.
id	72.685	39.749
d	0.335	0.472
t	-1.371	3.2
fpsa	0.834	4.006
tpsa	4.803	11.042
age	64.862	5.922
N	683	

```
                          ROC                    -Asymptotic Normal--
                   Obs    Area     Std. Err.     [95% Conf. Interval]
------------------------------------------------------------------------------
fpsa               683    0.7735     0.0187       0.73    0.81
tpsa               683    0.8375     0.0165       0.81    0.87
------------------------------------------------------------------------------
Ho: area(fpsa) = area(tpsa) chi2(1) =    40.17  Prob>chi2 =   0.0000
```

SAS – Example

The SAS code is, (I exported the data into SAS and created a work data set – temp)

```
proc logistic data=temp plots(only)=roc;;
  model d=y1 y2 / scale=none
                  clparm=wald
                  clodds=pl
                  rsquare;
 roc 'y1' y1;
 roc 'y2' y2;
   roccontrast reference('y1') / estimate e;
run;
```

Figure 2.1.3 contains the SAS output.

Problem – R Example

Use the CA cancer data and generate the ROC curve using R [R package - ROCR - can be used to create the ROC curve].

Diagnostic Likelihood Ratios

Another method of summarizing data of this type [albeit less commonly used] are the diagnostic likelihood ratios given by

$$\text{positive DLR} = \text{DLR}^{+} = \mathcal{LR}(y=1) \tag{2.17}$$

$$\text{negative DLR} = \text{DLR}^{-} = \mathcal{LR}(y=0) \tag{2.18}$$

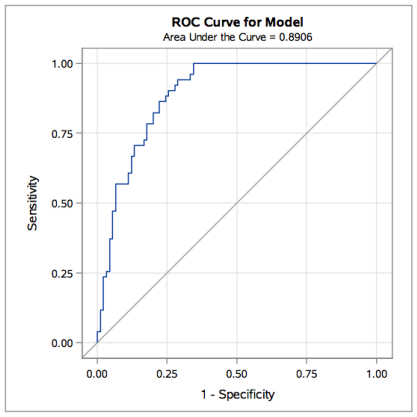
Pepe CA Cancer Data

The LOGISTIC Procedure

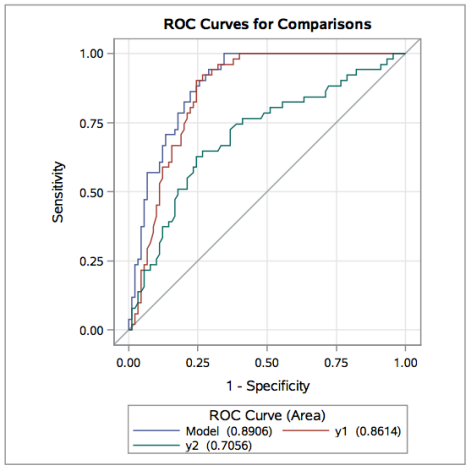
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.4645	0.3881	14.2422	0.0002
y1	1	-0.0274	0.00855	10.2803	0.0013
y2	1	-0.0163	0.00774	4.4133	0.0357

Parameter Estimates and Wald Confidence Intervals			
Parameter	Estimate	95% Confidence Limits	
Intercept	1.4645	0.7039	2.2251
y1	-0.0274	-0.0442	-0.0107
y2	-0.0163	-0.0314	-0.00109

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
y1	1.0000	0.973	0.954	0.986
y2	1.0000	0.984	0.967	0.998



The LOGISTIC Procedure



ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D (Gini)	Gamma	Tau-a
Model	0.8906	0.0263	0.8390	0.9422	0.7813	0.7813	0.3633
y1	0.8614	0.0306	0.8015	0.9214	0.7229	0.7241	0.3362
y2	0.7056	0.0468	0.6138	0.7973	0.4111	0.4123	0.1912

ROC Contrast Coefficients		
ROC Model	Row1	Row2
Model	1	0
y1	-1	-1
y2	0	1

where

$$\mathcal{LR}(y) \quad = \quad \frac{\Pr[Y = y \mid D = 1]}{\Pr[Y = y \mid D = 0]}. \tag{2.19}$$

Equation (2.19) is the ratio of the likelihoods for the observed test results in the diseased versus the non-diseased populations. An ideal test would have $\text{DLR}^+ = \infty$ and $\text{DLR}^- = 0$ whereas an uninformative test would have values $\text{DLR}^+ = \text{DLR}^- = 1$.

An attractive feature of these likelihood ratios is seen when comparing the pre-test odds with the post-test odds given by

$$\text{pre-test odds} \quad = \quad \frac{\Pr[D = 1]}{\Pr[D = 0]}$$

and

$$\text{post-test odds (Y)} \quad = \quad \frac{\Pr[D = 1 \mid Y]}{\Pr[D = 0 \mid Y]}$$

as

$$\begin{aligned} \text{post-test odds (Y=1)} &= \text{DLR}^+ \times \text{pre-test odds} && (2.20) \\ \text{post-test odds (Y=0)} &= \text{DLR}^- \times \text{pre-test odds.} && (2.21) \end{aligned}$$

Other results are,

$$\begin{aligned} \text{post-test odds (Y = 1)} &= \frac{\text{PPV}}{1 - \text{PPV}} \\ \text{post-test odds (Y = 0)} &= \frac{1 - \text{NPV}}{\text{NPV}} \\ \text{DLR}^+ &= \frac{\text{TPF}}{\text{FPF}} \\ \text{DLR}^- &= \frac{1 - \text{TPF}}{1 - \text{FPF}}. \end{aligned}$$

Estimation

Suppose that one tests N randomly selected subjects from a population with disease prevalence ρ where the results are summarized as,

	D = 0	D = 1	
Y = 0	$n_{\bar{D}}^-$	n_D^-	n^-
Y = 1	$n_{\bar{D}}^+$	n_D^+	n^+
	$n_{\bar{D}}$	n_D	N

Example

	D = 0	D = 1	
Y = 0	327	208	535
Y = 1	115	815	939
	442	1023	1465

$$
\begin{aligned}
N\hat{P}V &= 0.61 \\
P\hat{P}V &= 0.88 \\
F\hat{P}F &= 0.26 \\
T\hat{P}F &= 0.89
\end{aligned}
$$

Covid-19 Tests

I have included a diagram (Figure 2.4) taken from a paper that appeared in JAMA on May 6, 2020, some two months after the virus first appeared in this country. There are two types of tests displayed on this graph. The first is detection of the SARS-CoV-2 virus which infects the host about a week before the onset of symptoms. These tests are dependent upon swab PCR from differing locations and are represented by the solid lines. The tests for anti-bodies are serological ELISA based tests for two different antibodies and are represented by the dashed lines. In each case the detection probability is indicated by the height of the curve.

The authors indicate that the virus is most detectable within a week of symptom onset (within 2 weeks of infection) and is no longer detectable after two weeks from the onset time. Whereas the antibodies are detectable as soon as three weeks from onset and the IgG antibody remains detectable for a greater amount of time.

The accuracy and reliability of these tests are highly dependent upon when the test occurs relative to the disease onset and or the onset of symptoms. Suppose an individual is infected with no symptoms, what do you do? If one supposes that these people are infectious as long as the virus is detectable then they could be an unknown super spreader of the virus for nearly 30 days. What happens when you test someone who is not positive, which describes a very sizeable proportion of the US population, what happens to the reliability of the tests?

2.1.4 R × C Tables

Suppose that one has the following table,

Group	1	2	$\cdots$	c	Total
1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+c}	n

Tests of Independence for General Tables

The results given for the 2×2 tables can easily be extended to the general tables. The Pearson chi-square statistics becomes

$$
Q_P = \sum_{i}^{r} \sum_{j}^{c} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{2.22}
$$

where

$$
m_{ij} = E(n_{ij} \mid H_0) = \frac{n_{i+} n_{+j}}{n}.
$$

Figure. Estimated Variation Over Time in Diagnostic Tests for Detection of SARS-CoV-2 Infection Relative to Symptom Onset

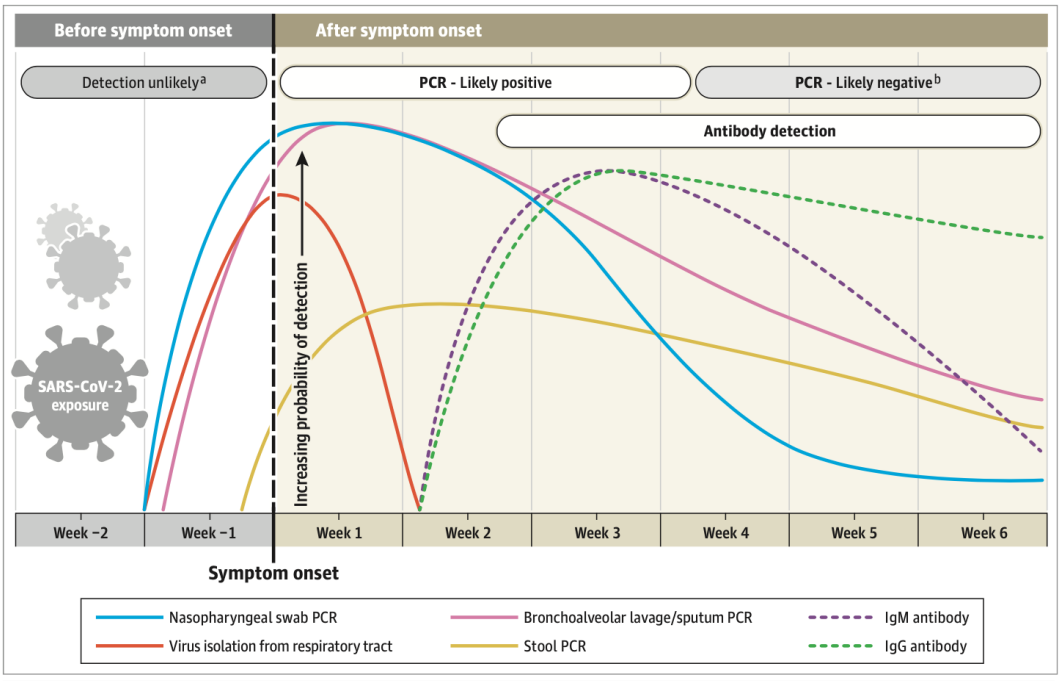


Figure 2.4: Covid-19 tests for SARS CoV-2

Q_P has an asymptotic chi-square distribution with $df = (r-1) \times (c-1)$.
As in the 2×2 case,

$$Q_C = \frac{n-1}{n} Q_P$$

has an asymptotic chi-square distribution with $df = (r-1) \times (c-1)$. This can be shown by observing that when one has fixed marginal totals the distribution of n_{ij} has a hypergeometric distribution under the null hypothesis of no association between the row and column random variables. That is,

$$\Pr[n_{ij}] = \frac{\prod_i^r n_{i+}! \prod_j^c n_{+j}!}{n! \prod_i^r \prod_j^c n_{ij}!}.$$

By[3] letting $n^* = (n_{11}, \ldots, n_{1c}, \ldots, n_{r1}, \ldots, n_{rc})'$ denote the $rc \times 1$ vector of observed frequencies, and let $m^* = E(n^*) = (m_{11}, \ldots, m_{1c}, \ldots, m_{r1}, \ldots, m_{rc})'$ denote the corresponding vector of expected frequencies. The expected value of n_{ij} is

$$m_{ij} = \frac{n_{i+} n_{+j}}{n} = n p_{i+} p_{+j}$$

where $p_{i+} = n_{i+}/n$ and $p_{+j} = n_{+j}/n$. Let $p_{*+} = (p_{1+}, \ldots, p_{r+})$ denote the $r \times 1$ vector of marginal row probabilities, and let $p_{+*} = (p_{+1}, \ldots, p_{+c})$ denote the $c \times 1$ vector of marginal column probabilities. It can be shown that

$$m^* = E(n^*) = n(p_{*+} \otimes p_{+*}),$$

where $\otimes$ is the direct or tensor or Kronecker product of matrices. The $rc \times rc$ covariance $Cov(n^*) = \Sigma$ has elements given by

$$\begin{aligned} cov(n_{ij}, n_{i'j'} \mid H_0) &= \frac{n_{i+}(n\delta_{ii'} - n_{i'+}) n_{+j}(n\delta_{jj'} - n_{+j'})}{n(n-1)} \\ &= \frac{n^2}{n-1} p_{i+}(\delta_{ii'} - p_{i'+}) p_{+j}(\delta_{jj'} - p_{+j'}) \end{aligned}$$

where $\delta_{kk'} = 1$ if $k = k'$ and is zero otherwise. In matrix notation Σ can be written as,

$$\Sigma = \frac{n^2}{n-1}(D_{p_{*+}} - p_{*+} p'_{*+}) \otimes (D_{p_{+*}} - p_{+*} p'_{+*})$$

where $D_{p_{*+}}$ and $D_{p_{+*}}$ are diagonal matrices with the elements p_{*+} and p_{+*}, on the main diagonal.
It can be shown that $n^* \sim N_{rc}(m^*, \Sigma)$ when the sample size n is large. Using the properties of quadratic forms of multivariate normal random variables it can be shown that

$$Q_C = (n^* - m^*)' A' (A \Sigma A')^{-1} A (n^* - m^*) = G' V_G^{-1} G,$$

has a chi-square distribution where A is given by

$$A = [I_{(r-1)}, 0_{(r-1)}] \otimes [I_{(c-1)}, 0_{(c-1)}]$$

and $G = A(n^* - m^*)$, $V_G = A \Sigma A'$.

[3] Note: the material that follows is based upon the properties of quadratic forms and the multivariate normal distribution. These materials were outlined in the first chapter. The actual details are topics that are usually covered in a multivariate analysis course.

Test of Association with q (r × c) Tables

Suppose that one has q independent tables and wishes to test the null hypothesis H_0 : no association between the row and column variables in any of the q tables. Let the h^{th} $r \times c$ table be given by

Group	1	2	$\cdots$	c	Total
1	n_{h11}	n_{h12}	$\cdots$	n_{h1c}	n_{h1+}
2	n_{h21}	n_{h22}	$\cdots$	n_{h2c}	n_{h2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	n_{hr1}	n_{hr2}	$\cdots$	n_{hrc}	n_{hr+}
Total	n_{h+1}	n_{h+2}	$\cdots$	n_{h+c}	n_h

Using the notation given in the previous section we have

$$\begin{aligned} m_h^* &= E(n_h^*) = n_h(p_{h*+} \otimes p_{h+*}), \\ \Sigma_h &= \frac{n_h^2}{n_h - 1}(D_{p_{h*+}} - p_{h*+}p'_{h*+}) \otimes (D_{p_{h+*}} - p_{h+*}p'_{h+*}) \end{aligned}$$

Define $G_h = A(n_h^* - m_h^*)$ and $G = \sum_{h=1}^q G_h$. It follows that $G \sim N_{(r-1)(c-1)}(0, V_G)$ where $V_G = \sum_{h=1}^q A\Sigma_h A'$, from which one can test this hypothesis with

$$Q_C = G'V_G^{-1}G \sim \chi^2(df = (r-1)(c-1)).$$

This statistics is known as the Cochran-Mantel-Haenszel (CMH) general association statistic.[4][5]

CMH Mean Score Test

Suppose that the set of q independent $r \times c$ tables are as above except that the column variable is ordinal. Furthermore, assume that scores, given by, $b_{h1}, \ldots, b_{hc}$ have been assigned to the levels of these column variables. The null hypothesis is H_0 : no association between the row and column variables in any of the q tables. Then

$$Q_M = M'V_M^{-1}M \sim \chi^2(df = r-1),$$

where $M_h = A_h(n_h^* - m_h^*)$, $A_h = (I_{r-1}, 0_{r-1}) \otimes (b_{h1}, \ldots, b_{hc})$, $M = \sum_{h=1}^q M_h$, and $V_M = \sum_{h=1}^q A_h\Sigma_h A'_h$.[6]

CMH Correlation Statistic

Suppose that the set of q independent $r \times c$ tables are as in the previous table. except that the row and column variables are both at least ordinal. Let the scores $a_{h1}, \ldots, a_{hr}$ and $b_{h1}, \ldots, b_{hc}$ be assigned to the levels of the row and column variables, respectively. The null hypothesis is H_0 : no association between the row and column variables in any of the q tables versus the alternative that there is a consistent positive (or negative) association between the row scores and the column scores across the tables. Then

$$Q_C = C'V_C^{-1}C = C^2/V_C \sim \chi^2(df = 1),$$

[4] Note: the degrees of freedom for the General Association Statistic is $df = (r-1)(c-1)$. The rows and columns are interchangeable in this analysis.

[5] **Caution**: The CMH statistics have low power for detecting an association in which the patterns of association for some of the strata are in the opposite direction of the patterns displayed by other strata. Thus, a nonsignificant CMH statistic suggests either that there is no association or that no pattern of association has enough strength or consistency to dominate any other pattern.

[6] Note: The degrees of freedom for the Mean Score Test is $df = (r-1)$. SAS requires that the columns be ordinal. One cannot interchange rows and columns for this analysis.

where $C_h = A_h(n_h^* - m_h^*)$, $A_h = (a_{h1}, \ldots, a_{hr}) \otimes (b_{h1}, \ldots, b_{hc})$, $C = \sum_{h=1}^{q} C_h$, and $V_C = \sum_{h=1}^{q} A_h \Sigma_h A_h'$. If $q = 1$, then $Q_C = (n-1)r^2$, where r^2 is the Pearson correlation between the row and column scores.[7]

2.1.5 Measures of Association

The following is from the SAS User's guide for PROC FREQ.

Tests and Measures of Agreement

When you specify the AGREE option in the TABLES statement, PROC FREQ computes tests and measures of agreement for square tables. For two-way tables, these tests and measures include McNemar's test for 2×2 tables, Bowker's test of symmetry, the simple kappa coefficient, and the weighted kappa coefficient. For multiple strata (n-way tables, where $n > 2$), PROC FREQ computes the overall simple kappa coefficient and the overall weighted kappa coefficient, as well as tests for equal kappas (simple and weighted) among strata. Cochran's Q is computed for multi-way tables when each variable has two levels, that is, for $2 \times 2 \times \ldots \times 2$ tables.

- **McNemar's Test** PROC FREQ computes McNemar's test for 2×2 tables when you specify the AGREE option. McNemar's test is appropriate when you are analyzing data from matched pairs of subjects with a dichotomous (yes-no) response. It tests the null hypothesis of marginal homogeneity, or $\pi_{1+} = \pi_{+1}$. McNemar's test is computed as

$$Q_M = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})} \sim \chi^2(df = 1).$$

- **Bowker's Test of Symmetry** For Bowker's test of symmetry, the null hypothesis is that the probabilities in the square table satisfy symmetry or that $\pi_{ij} = \pi_{ji}$ for all pairs of table cells. When there are more than two categories, Bowker's test of symmetry is calculated as

$$Q_B = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} \sim \chi^2(df = R(R-1)).$$

- **Simple Kappa Coefficient** The simple kappa coefficient, introduced by Cohen (1960), is a measure of interrater agreement:

$$\hat{\kappa} = \frac{P_o - P_e}{1 - P_e}$$

  where $P_o = \sum_i p_{ii}$ and $P_e = \sum_i p_{i+} p_{+i}$. If the two response variables are viewed as two independent ratings of the n subjects, the kappa coefficient equals +1 when there is complete agreement of the raters. When the observed agreement exceeds chance agreement, kappa is positive, with its magnitude reflecting the strength of agreement. Although this is unusual in practice, kappa is negative when the observed agreement is less than chance agreement. The minimum value of kappa is between -1 and 0, depending on the marginal proportions.

  The asymptotic variance of the simple kappa coefficient can be estimated by the following, according to Fleiss, Cohen, and Everitt (1969):

[7] Note: The degrees of freedom for the CMH Correlation Statistics is $df = 1$. SAS requires that both the rows and columns be ordinal. One can interchange rows and columns for this analysis.

$$var = \frac{(A + B - C)}{(n\ (1 - P_e)^2)}$$

where

$$A = \sum_i p_{ii}[1 - (p_{i+} + p_{+i})(1 - \hat{\kappa})]^2$$

$$B = (1 - \hat{\kappa})^2 \sum_{i \neq j} p_{ij}(p_{+i} + p_{i+})^2$$

and

$$C = \left[\hat{\kappa} - P_e(1 - \hat{\kappa})\right]^2.$$

PROC FREQ computes confidence limits for the simple kappa coefficient according to

$$\hat{\kappa} \pm z_{\alpha/2}\ \sqrt{var}$$

where $z_{\alpha/2}$ is the 100(1 - α/2) percentile of the standard normal distribution. The value of α is determined by the value of the ALPHA = option, which, by default, equals 0.05 and produces 95% confidence limits.

To compute an asymptotic test for the kappa coefficient, PROC FREQ uses a standardized test statistic $\hat{\kappa}^*$, which has an asymptotic standard normal distribution under the null hypothesis that kappa equals zero. The standardized test statistic is computed as

$$\hat{\kappa}^* = \hat{\kappa}/\sqrt{var_0(\hat{\kappa})}$$

where $var_0(\hat{\kappa})$ is the variance of the kappa coefficient under the null hypothesis.

$$var_0(\hat{\kappa}) = \frac{P_e + P_e^2 - \sum_i p_{i+}p_{+i}(p_{i+} + p_{+i})}{n\ (1 - P_e)^2}.$$

Refer to Fleiss (1981).

Ordinal Measures of Association

When you specify the MEASURES option in the TABLES statement, PROC FREQ computes several statistics that describe the association between the two variables of the contingency table. The following are measures of ordinal association that consider whether the variable Y tends to increase as X increases: gamma, Kendall's tau-b, Stuart's tau-c, and Somers' D. These measures are appropriate for ordinal variables, and they classify pairs of observations as concordant or discordant. A pair is concordant if the observation with the larger value of X also has the larger value of Y. A pair is discordant if the observation with the larger value of X has the smaller value of Y. Refer to Agresti (1996) and the other references cited in the discussion of each measure of association.

The Pearson correlation coefficient and the Spearman rank correlation coefficient are also appropriate for ordinal variables. The Pearson correlation describes the strength of the linear association between the row and column variables, and it is computed using the row and column scores specified by the SCORES= option in the TABLES statement. The Spearman correlation is computed with rank scores.

PROC FREQ computes estimates of the measures according to the formulas given in the discussion of each measure of association. For each measure, PROC FREQ computes an asymptotic standard error (ASE), which is the square root of the asymptotic variance denoted by var in the following sections.

Confidence Limits

If you specify the CL option in the TABLES statement, PROC FREQ computes asymptotic confidence limits for all MEASURES statistics. The confidence limits are computed as

$$est \pm z_{\alpha/2} \times ASE$$

where est is the estimate of the measure, $z_{\alpha/2}$ is the 100(1 - α/2) percentile of the standard normal distribution, and ASE is the asymptotic standard error of the estimate.

Asymptotic Tests

For each measure that you specify in the TEST statement, PROC FREQ computes an asymptotic test of the null hypothesis that the measure equals zero. To compute an asymptotic test, PROC FREQ uses a standardized test statistic z, which has an asymptotic standard normal distribution under the null hypothesis. The standardized test statistic is computed as

$$z = \frac{est}{\sqrt{var_0(est)}}$$

where est is the estimate of the measure and $var_0(est)$ is the variance of the estimate under the null hypothesis. Formulas for $var_0(est)$ are given in the discussion of each measure of association.

Note that the ratio of est to $var_0(est)$ is the same for the following measures: gamma, Kendall's tau-b, Stuart's tau-c, Somers' $D(R|C)$, and Somers' $D(C|R)$. Therefore, the tests for these measures are identical. For example, the p-values for the test of $H_0 : gamma = 0$ equal the p-values for the test of $H_0 : tau - b = 0$.

The available measures of association are:

- **Gamma** The estimator of gamma is based only on the number of concordant and discordant pairs of observations. It ignores tied pairs (that is, pairs of observations that have equal values of X or equal values of Y). Gamma is appropriate only when both variables lie on an ordinal scale. It has the range $-1 \leq \Gamma \leq 1$. If the two variables are independent, then the estimator of gamma tends to be close to zero. Gamma is estimated by

  $$G = \frac{(P - Q)}{(P + Q)}$$

  with asymptotic variance

  $$var = \frac{16}{(P+Q)^4} \sum_i \sum_j n_{ij} (Q\ A_{ij} - P\ D_{ij})^2$$

  where

  $$A_{ij} = \sum_{k>i} \sum_{l>j} n_{kl} + \sum_{k<i} \sum_{l<j} n_{kl}$$

  $$D_{ij} = \sum_{k>i} \sum_{l<j} n_{kl} + \sum_{k<i} \sum_{l>j} n_{kl}$$

$$P = \sum_i \sum_j n_{ij} A_{ij}$$

and

$$Q = \sum_i \sum_j n_{ij} D_{ij}.$$

The variance of the estimator under the null hypothesis that gamma equals zero is computed as

$$var_0(G) = \frac{4}{(P+Q)^2} \big(\sum_i \sum_j n_{ij} (A_{ij} - D_{ij})^2 - (P-Q)^2/n \big).$$

For 2×2 tables, gamma is equivalent to Yule's Q. Refer to Goodman and Kruskal (1979), Agresti (1990), and Brown and Benedetti (1977).

- **Kendall's Tau-b** Kendall's tau-b is similar to gamma except that tau-b uses a correction for ties. Tau-b is appropriate only when both variables lie on an ordinal scale. Tau-b has the range $-1 \leq \tau_b \leq 1$. It is estimated by

$$t_b = \frac{P-Q}{\sqrt{w_r \ w_c}}$$

with

$$var = w^{-4} \ \big(\sum_i \sum_j n_{ij} (2w d_{ij} + t_b \ v_{ij})^2 - n^3 t_b^2 (w_r + w_c)^2 \big)$$

where

$$w = \sqrt{w_r \ w_c}$$

$$w_r = n^2 - \sum_i n_{i+}^2$$

$$w_c = n^2 - \sum_j n_{+j}^2$$

$$d_{ij} = A_{ij} - D_{ij}$$

and

$$v_{ij} = n_{i+} \ w_c + n_{+j} \ w_r.$$

The variance of the estimator under the null hypothesis that tau-b equals zero is computed as

$$var_0(t_b) = \frac{4}{w_r \ w_c} \big(\sum_i \sum_j n_{ij} (A_{ij} - D_{ij})^2 - (P-Q)^2/n \big).$$

Refer to Kendall (1955) and Brown and Benedetti (1977).

- **Stuart's Tau-c** Stuart's tau-c makes an adjustment for table size in addition to a correction for ties. Tau-c is appropriate only when both variables lie on an ordinal scale. Tau-c has the range $-1 \leq \tau_c \leq 1$. It is estimated by

$$t_c = \frac{(m(P-Q))}{(n^2(m-1))}$$

with

$$var = \frac{4m^2}{(m-1)^2 n^4} \big(\sum_i \sum_j n_{ij} d_{ij}^2 - (P-Q)^2/n \big)$$

where $m = min(R, C)$.

The variance of the estimator under the null hypothesis that tau-c equals zero is

$$var_0(t_c) = var.$$

Refer to Brown and Benedetti (1977).

- **Somers' $D(C|R)$ and $D(R|C)$** Somers' $D(C|R)$ and Somers' $D(R|C)$ are asymmetric modifications of tau-b. $C|R$ denotes that the row variable X is regarded as an independent variable, while the column variable Y is regarded as dependent. Similarly, $R|C$ denotes that the column variable Y is regarded as an independent variable, while the row variable X is regarded as dependent. Somers' D differs from tau-b in that it uses a correction only for pairs that are tied on the independent variable. Somers' D is appropriate only when both variables lie on an ordinal scale. It has the range $-1 \leq D \leq 1$. Formulas for Somers' $D(R|C)$ are obtained by interchanging the indices.

$$D(C|R) = \frac{P - Q}{w_r}$$

with

$$var = \frac{4}{w_r^4}\big(\sum_i \sum_j n_{ij}(w_r\ d_{ij} - (P - Q)\ (n - n_{i+}))^2\big).$$

The variance of the estimator under the null hypothesis that $D(C|R)$ equals zero is computed as

$$var_0(D(C \mid R)) = \frac{4}{w_r^4}\big(\sum_i \sum_j n_{ij}(A_{ij} - D_{ij})^2 - (P - Q)^2/n\big).$$

Refer to Somers (1962), Goodman and Kruskal (1979), and Liebetrau (1983).

- **Pearson Correlation Coefficient** PROC FREQ computes the Pearson correlation coefficient using the scores specified in the SCORES= option. The Pearson correlation is appropriate only when both variables lie on an ordinal scale. It has the range $-1 \leq \rho \leq 1$. The Pearson correlation coefficient is computed as

$$r = \frac{v}{w} = \frac{ss_{rc}}{\sqrt{ss_r ss_c}}$$

with

$$var = w^{-4} \sum_i \sum_j n_{ij}(w(R_i - \bar{R})(C_j - \bar{C}) - \frac{b_{ij} v}{2w})^2$$

The row scores R_i and the column scores C_j are determined by the SCORES= option in the TABLES statement, and

$$\bar{R} = \sum_i n_{i+} R_i / n, \quad \bar{C} = \sum_j n_{+j} C_j / n$$

$$ss_r = \sum_i \sum_j n_{ij}(R_i - \bar{R})^2, \quad ss_c = \sum_i \sum_j n_{ij}(C_j - \bar{C})^2$$

$$ss_{rc} = \sum_i \sum_j n_{ij}(R_i - \bar{R})(C_j - \bar{C})$$

$$b_{ij} = (R_i - \bar{R})^2 ss_c + (C_j - \bar{C})^2 ss_r$$

and

$$v = ss_{rc}, \quad w = \sqrt{ss_r ss_c}.$$

Refer to Snedecor and Cochran (1989) and Brown and Benedetti (1977). To compute an asymptotic test for the Pearson correlation, PROC FREQ uses a standardized test statistic r^*, which has an asymptotic standard normal distribution under the null hypothesis that the correlation equals zero. The standardized test statistic is computed as

$$r^* = \frac{r}{\sqrt{var_0(r)}}$$

where $var_0(r)$ is the variance of the correlation under the null hypothesis.

$$var_0(r) = \frac{\sum_i \sum_j n_{ij}(R_i - \bar{R})^2(C_j - \bar{C})^2 - ss_{rc}^2/n}{ss_r ss_c}.$$

The asymptotic variance is derived for multinomial sampling in a contingency table framework, and it differs from the form obtained under the assumption that both variables are continuous and normally distributed. Refer to Brown and Benedetti (1977).

- **Spearman Rank Correlation** Coefficient The Spearman correlation coefficient is computed using rank scores $R1_i$ and $C1_j$, defined as,

$$R1_i = \sum_{k<i} n_{k+} + (n_{i+} + 1)/2 \quad C1_j = \sum_{l<j} n_{+l} + (n_{+j} + 1)/2.$$

It is appropriate only when both variables lie on an ordinal scale. It has the range $-1 \leq \rho_s \leq 1$. The Spearman correlation coefficient is computed as

$$r_s = \frac{v}{w}$$

with

$$var = \frac{1}{n^2 w^4} \sum_i \sum_j n_{ij}(z_{ij} - \bar{z})^2$$

where

$$v = \sum_i \sum_j n_{ij} R(i) C(j)$$

$$w = \frac{1}{12}\sqrt{FG}, \quad F = n^3 - \sum_i n_{i+}^3, \quad G = n^3 - \sum_j n_{+j}^3$$

$$R(i) = R1_i - n/2, \quad C(j) = C1_j - n/2$$

$$\bar{z} = \frac{1}{n} \sum_i \sum_j n_{ij} z_{ij}, \quad z_{ij} = w v_{ij} - v w_{ij}$$

$$v_{ij} = \frac{n}{2}\left[2R(i)C(j) + \sum_l n_{il} C(l) + \sum_k n_{kj} R(k) + 2\sum_l \sum_{k>i} n_{kl} C(l) + 2\sum_k \sum_{l>j} n_{kl} R(k)\right]$$

and

$$w_{ij} = \frac{-n}{96w}(F n_{+j}^2 + G n_{i+}^2)$$

Refer to Snedecor and Cochran (1989) and Brown and Benedetti (1977).

To compute an asymptotic test for the Spearman correlation, PROC FREQ uses a standardized test statistic r_s^*, which has an asymptotic standard normal distribution under the null hypothesis that the correlation equals zero. The standardized test statistic is computed as

$$r_s^* = \frac{r_s}{\sqrt{var_0(r_s)}}$$

where $var_0(r_s)$ is the variance of the correlation under the null hypothesis.

$$var_0(r_s) = \frac{1}{n^2 w^2} \sum_i \sum_j n_{ij}(v_{ij} - \bar{v})^2$$

where

$$\bar{v} = \sum_i \sum_j n_{ij} v_{ij}/n.$$

The asymptotic variance is derived for multinomial sampling in a contingency table framework, and it differs from the form obtained under the assumption that both variables are continuous and normally distributed. Refer to Brown and Benedetti (1977).

2.2 Continuous Variables

2.2.1 Measures of Association - Correlations

The commonly used correlation statistics are,

- **Pearson product-moment correlation** is a parametric measure of a linear relationship between two variables.
- **Spearman rank-order correlation** is a nonparametric procedure whereby the original data in Pearson's method are replaced by their ranks.
- **Kendall's tau-b coefficient** uses the number of concordances and discordances in paired observations.

Pearson Product-Moment

The Pearson product-moment correlation measures both the strength and the direction of a linear relationship[8] The population Pearson product-moment correlation, denoted ρ_{xy}, is

$$\rho_{xy} = \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{V}(x)\mathrm{V}(y)}} = \frac{\mathrm{E}(\,(x - \mathrm{E}(x))(y - \mathrm{E}(y))\,)}{\sqrt{\mathrm{E}(x - \mathrm{E}(x))^2\,\mathrm{E}(y - \mathrm{E}(y))^2}}$$

The sample Pearson product-moment correlation is

$$r_{xy} = \frac{ss_{xy}}{\sqrt{ss_{xx}ss_{yy}}}$$

where $\bar{x}$ is the sample mean of x, $\bar{y}$ is the sample mean of y and

$$ss_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}),\ ss_{xx} = \sum_i (x_i - \bar{x})^2,\ ss_{yy} = \sum_i (y_i - \bar{y})^2.$$

[8]If one variable X is an exact linear function of another variable Y, a positive relationship exists if the correlation is 1 and a negative relationship exists if the correlation is -1. If there is no linear predictability between the two variables, the correlation is 0. If the two variables are normal with a correlation 0, the two variables are independent. However, correlation does not imply causality because, in some cases, an underlying causal relationship might not exist.

Probability values for the Pearson correlation (let $r^2 = r_{xy}^2$) are computed by assuming

$$t \;=\; (n-2)^{1/2}\left(\frac{r^2}{1-r^2}\right)^{1/2} \sim \text{t-dist(df=n-2)}.$$

Spearman Rank-Order

Spearman rank-order correlation is a nonparametric measure of association based on the ranks of the data values and is given by,

$$\begin{aligned}\theta_{xy} &= \frac{\sum_i((R_i-\bar{R})(S_i-\bar{S})\,)}{\sqrt{\sum_i(R_i-\bar{R})^2\sum(S_i-\bar{S})^2}} \\ &= \frac{ss_{rs}}{\sqrt{ss_r\;ss_s}}.\end{aligned}$$

where R_i is the rank of x_i, S_i is the rank of y_i, $\bar{R}$ is the mean of the R_i values, and $\bar{S}$ is the mean of the S_i values. Probability values for the Spearman correlation (let $\theta^2 = \theta_{xy}^2$) are computed by assuming

$$t \;=\; (n-2)^{1/2}\left(\frac{\theta^2}{1-\theta^2}\right)^{1/2} \sim \text{t-dist(df=n-2)}.$$

Kendall's Tau-b

Kendall's tau-b is a nonparametric measure of association based on the number of concordances and discordances in paired observations. The Kendall's tau-b estimate of the correlation is

$$\tau_{xy} = \frac{\sum_{i<j}\left(\text{sgn}(x_i-x_j)\text{sgn}(y_i-y_j)\right)}{\sqrt{(T_0-T_1)(T_0-T_2)}}$$

where $T_0 = n(n-1)/2$, $T_1 = \sum_k t_k(t_k-1)/2$, and $T_2 = \sum_l u_l(u_l-1)/2$. The t_k is the number of tied x values in the k^{th} group of tied x values, u_l is the number of tied y values in the l^{th} group of tied y values, n is the number of observations, and $\text{sgn}(z)$ is defined as

$$\text{sgn}(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z = 0 \\ -1 & \text{if } z < 0 \end{cases}$$

Probability values for Kendall's tau-b are computed by assuming

$$\frac{s}{\sqrt{V(s)}} \sim N(0,1),$$

where

$$s = \sum_{i<j}\left(\text{sgn}(x_i-x_j)\text{sgn}(y_i-y_j)\right).$$

The variance of s is computed as

$$V(s) = \frac{v_0 - v_t - v_u}{18} + \frac{v_1}{2n(n-1)} + \frac{v_2}{9n(n-1)(n-2)}$$

where

$$\begin{aligned} v_0 &= n(n-1)(2n+5) \\ v_t &= \sum_k t_k(t_k-1)(2t_k+5) \\ v_u &= \sum_l u_l(u_l-1)(2u_l+5) \\ v_1 &= (\sum_k t_k(t_k-1))\,(\sum u_i(u_l-1)) \\ v_2 &= (\sum_l t_i(t_k-1)(t_k-2))\,(\sum u_l(u_l-1)(u_l-2)) \end{aligned}$$

The sums are over tied groups of values where t_i is the number of tied x values and u_i is the number of tied y values (Noether, 1967).

2.2.2 Test of Hypothesis for Two Populations

This section will contain a number of procedures for testing equality of two location and scale parameters using both parametric and non parametric procedures when the two samples are independent or dependent as in the paired design assumptions. The parametric method is covered in an introductory course, such as, STAT 2381. Since, you have seen this material before I will briefly cover the material and skip parts in the interest of time and the objectives of this course.

Independent Sample Design

The parametric method for testing hypotheses concerning the means for two normal populations is presented. Let $y_{1i} \sim N(\mu_1, \sigma_1^2)$ for $i = 1, \ldots, n_1$ denote a random sample from population 1 and $y_{2i} \sim N(\mu_2, \sigma_2^2)$ for $i = 1, \ldots, n_2$ denote a random sample from population 2 when σ_1, and σ_2 are unknown. The within group sample mean estimates ($\bar{y}_1$ and $\bar{y}_2$), sample standard deviation estimates (s_1 and s_2), standard errors (se_1 and se_2), and confidence limits for means and standard deviations are computed in the same way as for the one-sample design. The mean difference $\mu_1 - \mu_2 = \mu_d$ is estimated by $\bar{y}_d = \bar{y}_1 - \bar{y}_2$.

Under the assumption of **equal variances** ($\sigma_1^2 = \sigma_2^2$), the pooled estimate of the common standard deviation is

$$s_p = \left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \right)^{\frac{1}{2}}$$

The pooled standard error (the estimated standard deviation of $\bar{y}_d$ assuming equal variances) is

$$se_p = s_p \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{\frac{1}{2}}$$

The pooled $100(1-\alpha)\%$ confidence interval for the mean difference μ_d is

$$\left(\bar{y}_d \pm t_{1-\frac{\alpha}{2}, n_1+n_2-2} \times se_p \right)$$

The t value for the pooled test is computed as

$$t_p = \frac{\bar{y}_d - \mu_0}{se_p}$$

The two-sided p-value of the test is computed as

$$\Pr[t_p^2 > F_{1-\alpha,1,n_1+n_2-2}]$$

Under the assumption of **unequal variances** (called the the Behrens-Fisher problem), the un-pooled standard error is computed as

$$se_u = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{\frac{1}{2}}$$

Satterthwaite's (1946) approximation for the degrees of freedom is

$$df_u = \frac{se_u^4}{\frac{s_1^4}{(n_1-1)n_1^2} + \frac{s_2^4}{(n_2-1)n_2^2}}$$

The unpooled Satterthwaite $100(1-\alpha)\%$ confidence interval for the mean difference μ_d is

$$\left(\bar{y}_d \pm t_{1-\frac{\alpha}{2},df_u} se_u\right)$$

The t value for the unpooled Satterthwaite test is computed as

$$t_u = \frac{\bar{y}_d - \mu_0}{se_u}$$

The two-sided p-value of the unpooled Satterthwaite test is computed as

$$\Pr[t_u > F_{1-\alpha,1,df_u}]$$

When the COCHRAN option is specified in the PROC TTEST statement, the Cochran and Cox (1950) approximation of the p-value of the t_u statistic is the value of p such that

$$t_u = \frac{\left(\frac{s_1^2}{n_1}\right) t_1 + \left(\frac{s_2^2}{n_2}\right) t_2}{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}$$

where t_1 and t_2 are the critical values of the t distribution corresponding to a significance level of p and sample sizes of n_1 and n_2, respectively. The number of degrees of freedom is undefined when $n_1 \neq n_2$. In general, the Cochran and Cox test tends to be conservative (Lee and Gurland, 1975).

The $100(1-\alpha)\%$ CI=EQUAL and CI=UMPU confidence intervals for the common population standard deviation σ assuming equal variances are computed as discussed in the section Normal Data (DIST=NORMAL) for the one-sample design, except replacing s^2 by s_p^2 and $(n-1)$ by (n_1+n_2-1).

The folded form of the F statistic, F', tests the hypothesis that the variances are equal (Steel and Torrie, 1980), where

$$F' = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$$

The p-value gives the probability of a greater F value under the null hypothesis that $\sigma_1^2 = \sigma_2^2$.[9]

2.2.3 Paired Tests

The analysis is the same as the analysis for the one-sample design in the section Normal Data (DIST=NORMAL) based on the differences

$$d_i = y_{1i} - y_{2i}, \quad i \in \{1, \ldots, n\}.$$

[9]This test is not very robust to violations of the assumption that the data are normally distributed, and thus it is not recommended without confidence in the normality assumption.

2.2.4 Equivalence Tests

In this section a new wrinkle on the usual test of hypothesis. An alert should arise when your client is content when you can't reject the usual null hypothesis given by

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2.$$

This should send a red flag in terms of the analysis that was done. In reality, did the researcher NOT want to find significance? Were they trying to show that the groups were the same? When we don't reject, what can be said about the group means?

The Dilemma of the Non-rejected Null

- Fail to reject $\Rightarrow$ There is not statistically significant evidence that the population means are different. Without more information, we usually consider these groups similar. But this is under the idea that you're looking for evidence that they are different. If the researcher wants to claim that they're 'similar', it's not enough to do the traditional hypothesis test and just 'not reject'.
- Under our traditional hypothesis testing, we assume the null is true right from the start. (If you want to SHOW that the null is true, we certainly can't ASSUME it to be true.)
- If you want to show that the groups are similar, first ASSUME that they are different, and then try to gather evidence to the contrary (i.e. evidence that suggests they are the same).

  This is Equivalence Testing

  $$H_0 : \mu_1 \neq \mu_2 \text{ vs. } H_1 : \mu_1 = \mu_2.$$

- The difficult question in these tests...

  "How close is close enough to be considered 'the same'"?

- In equivalence testing, the null hypothesis is a "difference of Δ or more." Restating H_0 as

  $$H_0 : \mu_1 - \mu_2 < -\Delta \quad \text{or} \quad \mu_1 - \mu_2 > \Delta$$

- This leads to the most basic form of equivalence testing, the two one-sided test (TOST) procedure.

  $$\frac{(\bar{y}_1 - \bar{y}_2) + \Delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_{1-\alpha} \quad \text{or} \quad \frac{(\bar{y}_1 - \bar{y}_2) - \Delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -z_{1-\alpha}$$

  We declare the two group means equivalent at the α level if, and only if, both are rejected.

2.2.5 Simple Linear Rank Tests for Two-Sample Data

The material in this section may be new for you. It provides the needed theory for many of the non parametric methods for comparing parameters, such the median, from two populations that are not normally distributed.

Statistics of the form

$$S = \sum_{j=1}^{n} a(R_j)$$

are called *simple linear rank statistics*, where R_j is the rank of observation j, $a(R_j)$ is the score based on the rank of observation j, and n is the total number of observations[10].

[10] For two-sample data (where the observations are classified into two levels), PROC NPAR1WAY calculates simple linear rank statistics for the scores that you specify.

To compute an asymptotic test for a linear rank sum statistic, use the standardized test statistic z, which has an asymptotic standard normal distribution under the null hypothesis as

$$z = \frac{(S - E_0(S))}{\sqrt{Var_0(S)}}$$

where $E_0(S)$ is the expected value of S under the null hypothesis, and $Var_0(S)$ is the variance under the null hypothesis. As shown in Randles and Wolfe (1979),

$$E_0(S) = \frac{n_1}{n} \sum_{j=1}^{n} a(R_j)$$

where n_1 is the number of observations in the first class level (sample), n_2 is the number of observations in the other class level, and

$$Var_0(S) = \frac{n_1 n_2}{n(n-1)} \sum_{j=1}^{n} (a(R_j) - \bar{a})^2$$

where $\bar{a}$ is the average score,

$$\bar{a} = \frac{1}{n} \sum_{j=1}^{n} a(R_j)$$

2.2.6 Scores for Linear Rank Tests

The following score types are used primarily to test for differences in:

- Location
  - Wilcoxon, median, Van der Waerden (normal), Savage, and Conover.
- Scale
  - Siegel-Tukey, Ansari-Bradley, Klotz, Mood, and Conover.

Conover scores can be used to test for differences in both location and scale. This section gives formulas for the score types. For further information about the formulas and the applicability of each score, see Randles and Wolfe (1979), Gibbons and Chakraborti (2010), Conover (1999), and Hollander and Wolfe (1999).

Wilcoxon Scores

Wilcoxon scores are the ranks of the observations, $a(R_j) = R_j$ where R_j is the rank of observation j. The Wilcoxon scores in the linear rank statistic for two-sample data are used to perform the rank sum statistic for the Mann-Whitney-Wilcoxon test. The Wilcoxon scores are used in the one-way ANOVA statistic to perform the Kruskal-Wallis test.

Median Scores

Median scores equal 1 for observations greater than the median, and 0 otherwise. In terms of the observation ranks, median scores are defined as

$$a(R_j) = \begin{cases} 1 & \text{if} \quad R_j > (n+1)/2 \\ 0 & \text{if} \quad R_j \leq (n+1)/2 \end{cases}$$

Use the median scores in the linear rank statistic for two-sample data to produce the two-sample median test. The one-way ANOVA statistic with median scores is equivalent to the Brown-Mood test. Median scores are particularly powerful for distributions that are symmetric and heavy-tailed.

Van der Waerden (Normal) Scores

Van der Waerden scores are the quantiles of a standard normal distribution and are also known as *quantile normal* scores. Van der Waerden scores are computed as

$$a(R_j) = \Phi^{-1}\left(\frac{R_j}{n+1}\right)$$

where Φ is the cumulative distribution function of a standard normal distribution. These scores are powerful for normal distributions.

Savage Scores

Savage scores are expected values of order statistics from the exponential distribution, with 1 subtracted to center the scores around 0. Savage scores are computed as

$$a(R_j) = \sum_{i=1}^{R_j}\left(\frac{1}{n-i+1}\right) - 1$$

Savage scores are powerful for comparing scale differences in exponential distributions or location shifts in extreme value distributions (Hajek, 1969, p. 83).

Siegel-Tukey Scores

$$\begin{array}{llll} a(1)=1, & a(n)=2, & a(n-1)=3, & a(2)=4, \\ a(3)=5, & a(n-2)=6, & a(n-3)=7, & a(4)=8, \ \ldots \end{array}$$

where the score values continue to increase in this pattern toward the middle ranks until all observations have been assigned a score.

Ansari-Bradley Scores

Ansari-Bradley scores are similar to Siegel-Tukey scores, but Ansari-Bradley scoring assigns the same score value to corresponding extreme ranks. The Siegel-Tukey scores are a permutation of the ranks $1, 2, \ldots, n$. Ansari-Bradley scores are defined as

$$\begin{array}{ll} a(1)=1, & a(n)=1, \\ a(2)=2, & a(n-1)=2, \ \ldots \end{array}$$

Equivalently, Ansari-Bradley scores are equal to

$$a(R_j) = \frac{n+1}{2} \ - \ \left|R_j - \frac{n+1}{2}\right|$$

Klotz Scores

Klotz scores are the squares of the Van der Waerden (normal) scores. Klotz scores are computed as

$$a(R_j) = \left(\Phi^{-1}\left(\frac{R_j}{n+1}\right)\right)^2$$

where Φ is the cumulative distribution function of a standard normal distribution.

Mood Scores

Mood scores are computed as the square of the difference between the observation rank and the average rank. Mood scores can be written as

$$a(R_j) = \left(R_j - \frac{n+1}{2} \right)^2$$

Conover Scores

Conover scores are based on the squared ranks of the absolute deviations from the sample means. For observation j the absolute deviation from the mean is computed as

$$U_j = |X_{j(i)} - \bar{X}_i|$$

where $X_{j(i)}$ is the value of observation j, observation j belongs to sample i, and $\bar{X}_i$ is the mean of sample i. The values of U_j are ranked, and the Conover score for observation j is computed as

$$\text{Score}_j = (\text{Rank}(U_j))^2$$

The Conover score test is also known as the squared ranks test for variances. See Conover (1999) for more information.

2.2.7 Wilcoxon and Mann-Whitney Test

The Mann-Whitney and Wilcoxon test assumes that

- The data consists of a random sample of n_1 values, denoted $X_1, X_2, \ldots, X_{n_1}$ from $F_X(x)$ with $median(X) = \theta_X$, and a random sample of n_2 values, denoted $Y_1, Y_2, \ldots, Y_{n_2}$ from $F_Y(y)$ with $median(Y) = \theta_Y$.
- The random samples are at least ordinal.
- F_X and F_Y differ only with respect to the their median. That is, $\theta_X = \theta_Y + \delta$.

The null hypothesis is $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ (the usual one-sided tests are possible). The procedure is

1. Combine the data for the two random samples and rank the combined data where $r_j = rank(Y_j)$.
2. Compute the Wilcoxon statistics as, $W = \sum_{j=1}^{n_2} r_j$.
3. Reject H_0 if W is either too small ($\theta_Y < \theta_X$) or too large ($\theta_Y > \theta_X$).
4. The large sample distribution of $W^* = \frac{W - E(W)}{\sqrt{Var(W)}} \sim N(0,1)$, where

$$E(W) = \frac{n_2(N+1)}{2}$$

and

$$Var(W) = \frac{n_1 n_2 (N+1)}{12}$$

for $N = n_1 + n_2$.

The Mann-Whitney test statistics is

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(X_i, Y_j)$$

where $\phi(x, y) = I_{x<y}$ the indicator function. The two statistics are similar in that

$$W = U + \frac{n_2(n_2+1)}{2}.$$

Note: $\Pr[X < Y] = E(I_{X<Y})$ in which case the statistic U can be used as a nonparametric estimate of $\Pr[X < Y]$.[11]

2.2.8 Fligner-Policello Test

The Fligner-Policello (FP) location test[12] for two-sample data is used to test the null hypothesis $H_0 : \theta_X = \theta_Y$, where θ_X and θ_Y are the population medians of the two classes.[13] The (FP) test is based on placement scores (Orban and Wolfe, 1979) where the placement of an observation X_i from population 1, $P(X_i)$, is defined as the number of observations in the sample from population 2 that are less than X_i. If there are ties, the placement of X_i is adjusted by adding half the number of observations Y in the second sample that are equal to X_i. The placement of an observation Y_j from the second sample, $P(Y_j)$, is computed in the same way. The placement scores are,

$$Pl(X_i) = \sum_{j=1}^{n_y} (\ I(Y_j < X_i) \ + \ 0.5I(Y_j = X_i) \ )$$

$$Pl(Y_j) = \sum_{i=1}^{n_x} (\ I(X_i < Y_j) \ + \ 0.5I(X_i = Y_j) \ )$$

where $I(\cdot)$ is an indicator function and n_x and n_y denote the number of observations in two samples, respectively.

The Fligner-Policello test statistic is computed as

$$z = \left(\sum_{j=1}^{n_y} Pl(Y_j) - \sum_{i=1}^{n_x} Pl(X_i) \right) \ / \ \left(2\sqrt{V_x + V_y + \bar{Pl}_x \bar{Pl}_y} \right)$$

where

$$\bar{Pl}_x = \left(\sum_{i=1}^{n_x} Pl(X_i) \right) / n_x \quad V_x = \sum_{i=1}^{n_x} (Pl(X_i) - \bar{Pl}_x)^2$$

$$\bar{Pl}_y = \left(\sum_{j=1}^{n_y} Pl(Y_j) \right) / n_y \quad V_y = \sum_{j=1}^{n_y} (Pl(Y_j) - \bar{Pl}_y)^2.$$

Under the null hypothesis, the (FP) statistic has an asymptotic standard normal distribution.

[11] Since $\Pr[X < Y]$ = AUC, the Mann-Whitney statistics is a nonparametric estimate for the area under the ROC curve (AUC). Delong presented a method for computing the standard error for the Mann-whitney statistic.

[12] (Fligner and Policello, 1981) [FP option, PROC NPAR1WAY]

[13] The Fligner-Policello test is valid when $F_X(x)$ and $F_Y(y)$ are symmetric about the median, but it does not require that the two distributions have the same form or that the variances be equal. See Hollander and Wolfe (1999) and Juneau (2007) for more information.

2.2.9 Tests Based on the Empirical Distribution Function (EDF)

This section describes three nonparametric tests that are based on the empirical distribution function[14]. The procedures are; the Kolmogorov-Smirnov and Cramer-von Mises tests, and also the Kuiper test for two-sample data.[15] The null hypothesis is $H_0 : F_X(\cdot) = F_Y(\cdot) = F(\cdot)$. The (EDF) of a sample $\{x_j\}$, $j = 1, 2, \ldots, n$ is defined as

$$\hat{F}(x) = \frac{1}{n}(\text{number of } x_j \leq x) = \frac{1}{n}\sum_{j=1}^{n} I(x_j \leq x)$$

where $I(\cdot)$ is an indicator function. Let $\hat{F}_i$ denote the sample EDF for the i^{th} group. The EDF for the overall sample, pooled over groups, can also be expressed as

$$\hat{F}(x) = \frac{1}{n}\sum_{i}\left(n_i\hat{F}_i(x)\right)$$

where n_i is the number of observations in the i^{th} group, and n is the total number of observations.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov statistic measures the maximum deviation of the EDF within the groups from the pooled EDF. The Kolmogorov-Smirnov statistic is computed as,

$$KS = \max_j \sqrt{\frac{1}{n}\sum_i n_i\left(\hat{F}_i(x_j) - \hat{F}(x_j)\right)^2} \quad \text{for} \quad j = 1, 2, \ldots, n$$

The asymptotic Kolmogorov-Smirnov statistic is computed as, $KS_a = KS \times \sqrt{n}$. If there are only two class levels, the two-sample Kolmogorov-Smirnov test statistic D is

$$D = \max_j \left|\hat{F}_1(x_j) - \hat{F}_2(x_j)\right| \quad \text{for} \quad j = 1, 2, \ldots, n$$

The p-value for this test is the probability that D is greater than the observed value d under the null hypothesis of no difference between class levels (samples). The asymptotic p-value for D is approximated as,

$$\Pr(D > d) = 2\sum_{i=1}^{\infty}(-1)^{(i-1)}\ e^{(-2i^2z^2)}$$

where

$$z = d\sqrt{n_1 n_2/n}$$

See Hodges (1957) for information about this approximation.

Cramer-von Mises Test

The Cramer-von Mises statistic is

$$CM = \frac{1}{n^2}\sum_i\left(n_i\sum_{j=1}^{p} t_j\left(\hat{F}_i(x_j) - \hat{F}(x_j)\right)^2\right)$$

[14] [PROC NPAR1WAY - EDF option].

[15] For further information about the formulas and the interpretation of EDF statistics, see Hollander and Wolfe (1999) and Gibbons and Chakraborti (2010). For details about the k-sample analogs of the Kolmogorov-Smirnov and Cramer-von Mises statistics, see Kiefer (1959).

where t_j is the number of ties at the j^{th} distinct value and p is the number of distinct values. The asymptotic value is computed as

$$CM_a = CM \times n.$$

Kuiper Test

For data with two class levels, the Kuiper statistic is

$$K = \max_j \left(\hat{F}_1(x_j) - \hat{F}_2(x_j)\right) - \min_j \left(\hat{F}_1(x_j) - \hat{F}_2(x_j)\right) \quad \text{where} \quad j = 1, 2, \ldots, n$$

The asymptotic value is

$$K_a = K\sqrt{n_1 n_2 / n}$$

The p-value for the Kuiper test is the probability of observing a larger value of K_a under the null hypothesis of no difference between the two classes Owen (1962, p 441).

Part II

Regression Models

Chapter 3

General Linear Regression Model

Note: Since you have already seen simple linear regression in STAT 2381 and STAT 3386, I will briefly cover the material using a matrix notation.

The general linear[1] regression model can be written as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{p-1} x_{(p-1)i} + \epsilon_i \tag{3.1}$$

for $i = 1, 2, \ldots, n$ where the independent variables $x_{1i}, x_{2i}, \ldots, x_{(p-1)i}$ satisfy one of the following:

1. **Polynomial Regression** – when $x_{ji} = x_i^j, \quad i = 1, 2, \ldots, n, j = 1, 2, \ldots, (p-1)$. Equation (3.1) is a polynomial of degree $(p-1)$.
2. **Multiple Regression** – where each independent variable is distinct.

Equation (3.1) could be a combination of the above two models, however, in this chapter we will consider the model to be either polynomial or multiple regression. As in the previous chapter, the unobserved error ϵ_i is the vertical distance that the observed y_i is from the curve or surface given by $E(y_i \mid X\beta) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{p-1} x_{(p-1)i}$. Equation (3.1) can be written as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \ldots & x_{(p-1)1} \\ 1 & x_{12} & x_{22} & \ldots & x_{(p-1)2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{(p-1)n} \end{pmatrix} = \begin{pmatrix} \mathbf{j_n} & \mathbf{X}_* \end{pmatrix} \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

and

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix},$$

where

[1] Linear means that the expected value of the dependent variable can be expressed as an additive model in terms of the independent variables, versus having multiplicative or nonlinear terms.

$\mathbf{y}$ is a $n \times 1$ vector of dependent observations.

$\mathbf{X}$ is a $n \times p$ matrix of independent observations.

$\mathbf{x_j}$ is a $n \times 1$ vector for the j^{th} independent variable, $j = 1, 2, \ldots, p-1$.

$\mathbf{X}_* = \begin{pmatrix} \mathbf{x_1} & \mathbf{x_2} & \cdots & \mathbf{x_{p-1}} \end{pmatrix}$.

β is a $p \times 1$ vector of parameters.

ϵ is a $n \times 1$ vector of unobserved errors.

As in the previous chapter the least squares estimate for β is found by minimizing

$$Q(\beta) = \epsilon'\epsilon.$$

The solution to the minimization problem satisfies

$$\frac{\partial Q(\beta)}{\partial \beta} = -2X'y + 2X'X\beta = 0.$$

The above can be written as the normal equations

$$X'X\beta = X'y.$$

If $rank[X] = p$, [X is said to be full column rank], the normal equations have an unique solution given by

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

Under what conditions would the rank of X be less than p? Can this happen in polynomial regression? Can this be avoided, assuming that the choice of the $x_i's$ is yours? What about in multiple regression? What does having less than full column rank (rank of $X < p$) mean in multiple regression?

3.1 Inference

Let $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, 2, \ldots, n$, in which case we have

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \sim N_n(0, \sigma^2 I_n)$$

and

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \sim N_n(X\beta, \sigma^2 I_n).$$

Since, $\hat{\beta} = Ly$, $\hat{y} = Hy$, and $\hat{\epsilon} = (I - H)y$ for $L = (X'X)^{-1}X'$ and $H = X(X'X)^{-1}X'$, one has

1. $\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$.
   (a) $\hat{\beta}$ is an unbiased estimate of β.

(b) $var(\hat{\beta}_i) = \sigma^2((X'X)^{-1})_{ii}$.

(c) $cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2((X'X)^{-1})_{ij}$.

(d) $corr(\hat{\beta}_i, \hat{\beta}_j) = ((X'X)^{-1})_{ij}/[((X'X)^{-1})_{ii}((X'X)^{-1})_{jj}]^{1/2}$.

2. $\hat{y} \sim N_n(X\beta, \sigma^2 H)$.

(a) $var(\hat{y}_i) = \sigma^2 h_{ii}$.

(b) $cov(\hat{y}_i, \hat{y}_j) = \sigma^2 h_{ij}$, where $H = (h_{ij})$. Notice that the $\hat{y}_i's$ are not independent of one another unless each of the $h_{ij} = 0$.

(c) $corr(\hat{y}_i, \hat{y}_j) = h_{ij}/[h_{ii}h_{jj}]^{1/2}$.

3. $\hat{\epsilon} \sim N_n(0, \sigma^2(I - H))$.

(a) $var(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$.

(b) $cov(\hat{\epsilon}_i, \hat{\epsilon}_j) = -\sigma^2 h_{ij}$.

(c) $corr(\hat{\epsilon}_i, \hat{\epsilon}_j) = -h_{ij}/[(1 - h_{ii})(1 - h_{jj})]^{1/2}$.

3.1.1 Estimation of σ^2

The estimation of σ^2 follows from observing that

$$\begin{aligned} E(Q(\hat{\beta}) = y'(I - H)y) &= \sigma^2 tr[(I - H)] + \beta' X'(I - H)X\beta \\ &= \sigma^2 tr[(I - H)] \\ &= \sigma^2(tr[I] - tr[H]) \\ &= \sigma^2(n - p), \end{aligned}$$

where $E(y) = X\beta$ and $cov(y) = V = \sigma^2 I_n$. In which case, the least squares estimate for σ^2 is

$$\hat{\sigma}^2 = y'(I - H)y/(n - p) = SS_E/(n - p).$$

3.1.2 ANOVA Table

As in the previous chapter one has the analysis of variance table given by

Source	Sum of Squares	Degrees of Freedom	Mean Square
due to β	$SS(\beta) = \hat{\beta}' X'y = y'Hy$	p	$MS(\beta) = SS(\beta)/p$
Residual	$SS_E = y'y - \hat{\beta}' X'y = y'(I - H)y$	n-p	$MS_E = SS_E/(n - p)$
Uncorrected Total	$y'y$	n	

Since β_0 is a free parameter, the sum of squares term is adjusted for β_0, that is

$$y'Hy = y'(H - \frac{1}{n}\mathbf{j}\mathbf{j}')y + \frac{1}{n}y'\mathbf{j}\mathbf{j}'y$$

or

$$SS(\beta) = SS(\beta_0, \beta_*) = SS(\beta_* \mid \beta_0) + SS(\beta_0),$$

where $SS(\beta_0) = \frac{1}{n}y'\mathbf{j}\mathbf{j}'y = n\bar{y}^2$ is the correction factor or the sum of squares due to β_0. In which case the ANOVA table become

Source	Sum of Squares	Degrees of Freedom	Mean Square
due to $\beta_* \mid \beta_0$	$SS(\beta_* \mid \beta_0) = y'Hy - \frac{1}{n}y\mathbf{jj}'y$	p-1	$MS(\beta_* \mid \beta_0) = SS(\beta_* \mid \beta_0)/(p-1)$
Residual	$SS_E = y'y - \hat{\beta}'X'y = y'(I-H)y$	n-p	$MS_E = SS_E/(n-p)$
Corrected Total	$SS_{CT} = y'y - \frac{1}{n}y'\mathbf{jj}'y$	n-1	
due to β_0	$SS(\beta_0) = \frac{1}{n}y'\mathbf{jj}'y = n\bar{y}^2$	1	
Uncorrected Total	$y'y$	n	

where $\beta_* = (\beta_1, \beta_2, \ldots, \beta_{p-1})$

3.1.3 Expected Values of the Sum of Squares

As before, it follows when $V = \sigma^2 I_n$ that

1. $E(y'Hy) = \sigma^2 tr[H] + \beta'X'HX\beta = p\sigma^2 + \beta'X'X\beta$.
2. $E(y'(I-H)y) = \sigma^2 tr[(I-H)] + \beta'X'(I-H)X\beta = (n-p)\sigma^2$.

The R^2 is an indicator of how much of the variation in the data is explained by the model. It is defined as

$$R^2 = \frac{SS(\beta_* \mid \beta_0)}{SS_{CT}} = \frac{y'Hy - \frac{1}{n}y'\mathbf{jj}'y}{y'y - \frac{1}{n}y'\mathbf{jj}'y}.$$

The adjusted R^2 is the R^2 value adjusted for the number of parameters in the model and is given by

$$adjR^2 = 1 - [(n-i)/(n-p)](1-R^2)$$

where i is 1 if the model includes the y intercept (β_0), and is 0 otherwise.

3.2 Distribution of the Mean Squares

Again from the properties of the distribution of quadratic forms it can be show that

1. $SS(\beta)/\sigma^2 \sim \chi^2(df = p, \lambda = 1/2\beta'X'X\beta)$.
2. $SS_E/\sigma^2 \sim \chi^2(df = n-p)$.
3. $SS(\beta)$ and SS_E are independent.
4. $F = MS(\beta)/MS_E \sim F(df_1 = p, df_2 = n-p), \lambda = 1/2\beta'X'X\beta)$.
5. When $\beta = 0$ it follows that $SS(\beta)/\sigma^2 \sim \chi^2(df = p)$ and $F = MS(\beta)/MS_E \sim F(df_1 = p, df_2 = n-p)$.
6. $SS(\beta_* \mid \beta_0)/\sigma^2 \sim \chi^2(df = p-1, \lambda = 1/2\beta_*'X_*'X_*\beta_*)$, where

$$X_* = \begin{pmatrix} x_{11} & x_{21} & \ldots & x_{(p-1)1} \\ x_{12} & x_{21} & \ldots & x_{(p-1)1} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \ldots & x_{(p-1)n} \end{pmatrix}$$

and

$$\beta_* = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}.$$

7. $SS(\beta_* \mid \beta_0)$ and SS_E are independent.
8. $F = MS(\beta_* \mid \beta_0)/MS_E \sim F(df_1 = p-1, df_2 = n-p, \lambda = 1/2\beta_*' X_*' X_* \beta_*)$.
9. When $\beta_* = 0$ it follows that $SS(\beta_* \mid \beta_0)/\sigma^2 \sim \chi^2(df = p-1)$ and $F = MS(\beta_* \mid \beta_0)/MS_E \sim F(df_1 = p-1, df_2 = n-p)$.
10. In the general linear regression model, one rejects the null hypothesis

$$H_0 : \beta_* = 0 \quad \text{versus} \quad H_1 : \beta_* \neq 0$$

if $F = MS(\beta_* \mid \beta_0)/MS_E > F_\alpha(p-1, n-p)$. Then the power is computed as

$$\text{Power}(\beta_*) = \Pr[W > F_\alpha(p-1, n-p)].$$

where $W \sim F(p-1, n-p, \lambda = 1/2\beta_*' X_*' X_* \beta_*)$.

3.2.1 The Reduction Notation

In constructing the ANOVA tables one is interested in describing the sum of squares attributed to various terms in the model. This can be done in one of two ways, sequential or partial. Suppose that one has the three term model given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i.$$

If the model is a polynomial regression model of order 2 (i.e., $x_{ji} = x_i^j$) then the model sum of squares SS_M should be written as (sequential ordering)

$$SS_M = S(\beta_0) + S(\beta_1 \mid \beta_0) + S(\beta_2 \mid \beta_0, \beta_1) + S(\beta_3 \mid \beta_0, \beta_1, \beta_2).$$

Where each term is an independent quadratic forms with non-central chi-square distributions with degrees of freedom = 1.

When the model is a multiple regression model, one is interested in determining the amount of reduction that can be attributed to each variable (partial ordering)

$$S(\beta_3 \mid \beta_0, \beta_1, \beta_2)$$
$$S(\beta_2 \mid \beta_0, \beta_1, \beta_3)$$
$$S(\beta_1 \mid \beta_0, \beta_2, \beta_3).$$

These quadratic forms are no longer independent and their sum does not equal SS_M. However, they are independent of the error sum of squares and they can be shown to be a non-central chi-square distributed with one degree of freedom.

When using SAS the sequential sum of squares is the standard output and can be specified with the Type I SS statement. The partial SS can be obtained using the Type II SS statement. The Type I SS should be used when

- Balance ANOVA model with terms in their proper sequence or order.
- Purely nested model in proper order.
- **Polynomial regression** models.

The Type II SS statement should be used when

- The model is balanced.
- Any main effects model.
- **Multiple regression** model.
- An effect not contained in any other effect (no nesting).

3.2.2 Testing Linear Hypothesis

The general form of a linear hypothesis for the parameters is

$$H_0 : L\beta = c$$

where L is a $q \times p$ matrix of rank q. The approach is to estimate $L\beta - c$ with $L\hat{\beta} - c$ where

- $E(L\hat{\beta} - c) = L\beta - c$.
- $Cov(L\hat{\beta} - c) = \sigma^2 L(X'X)^{-1}L'$.
- $Q/\sigma^2 = (L\hat{\beta} - c)'(L(X'X)^{-1}L')^{-1}(L\hat{\beta} - c) \sim \chi^2(df = q, \lambda = 1/2(L\beta - c)'(L(X'X)^{-1}L')^{-1}(L\beta - c))$.
- $Q/q/\hat{\sigma}^2 \sim F(q, n - p)$ whenever $L\beta = c$.

Chapter 4

Checking Model Assumptions

In this chapter the validity of the assumptions given in the previous chapter are examined. Various techniques have been proposed for verifying whether or not the assumptions are valid.

4.1 Checks for Normality

4.1.1 Visual plots for the Residuals

If the proposed model assumption hold then the residuals satisfy $\hat{e}_i \sim N(0, \sigma^2(1-h_{ii}))$. A simple plots of the residuals versus the predicted values $\hat{y}_i$ or the j^{th} independent variable x_{ji} often reveals if this assumption has been satisfied. Draper and Smith has an extensive discussion concerning plots of this type. Essentially, you should not "see" any patterns in the scatterplots. Likewise, the amount of variability should be somewhat constant across the plots.

4.1.2 QQ and PP Plots

The QQ or PP plots can be used to assess normality of the residuals. That is, let $z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)}$ represent the ordered values of n independent and identically distributed $N(0,1)$ random variables. It can be shown that the expected value of $z_{(i)}$ is

$$E(z_{(i)}) \approx \gamma_i = \Phi^{-1}[(i - 3/8)/(n + 1/4)]$$

where Φ is the cdf for the standard normal given by

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-1/2t^2} dt.$$

The QQ plot consists of a scatterplot of $(z_{(i)}, \gamma_i)$. If the data are normal then the resulting scatterplot should lie close to the line $(\gamma_i = z_{(i)})$. The PP plot is similar to the QQ plot using the ordered pairs $(\Phi(z_{(i)}), [i/n])$. If either plot differs greatly from the diagonal line, then the normality assumption likely does not hold. Note, one should only check these plots when the model is appropriate as the residuals for inappropriate models often appear to be non-normally distributed.

4.2 Residual Analysis

4.2.1 Standardized Residuals

From the previous chapter we have that $\hat{e}_i = y_i - \hat{y}_i \sim N(0, \sigma^2(1 - h_{ii}))$. Since σ^2 is unknown it can be estimated with either

$$\hat{\sigma}^2 = MS_E$$

or

$$\hat{\sigma}^2_{(i)} = \frac{(n-p)\hat{\sigma}^2 - \hat{e}_i^2/(1-h_{ii})}{(n-p-1)}$$

where $\hat{\sigma}^2_{(i)}$ is the mean square for the error whenever the i^{th} observation has been omitted from the regression model. SAS provides two standardized residuals. They are:

1. Internally Studentized Residual (STUDENT) is given by

$$s_i = \frac{\hat{e}_i}{\hat{\sigma}(1-h_{ii})^{1/2}}.$$

2. Externally Studentized Residual (RSTUDENT) is given by

$$s_{(i)} = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}(1-h_{ii})^{1/2}}.$$

3. Their properties are;
   - $s_i \sim t(df = n-p)$.
   - $s_{(i)} \sim t(df = n-p-1)$.

4.2.2 Leverages

The hat matrix H has the following properties;

1. $SS(\beta) = \hat{\beta}'X'y = y'Hy = y'H^2y = \hat{y}'\hat{y}$.
2. $\sum_{i=1}^n var(\hat{y}_i)/n = tr[\sigma^2 H]/n = \sigma^2 p/n$.
3. $H\mathbf{j} = \mathbf{j}$ whenever the y intercept is included in the model. In which case, the sum of every row and every column of H equals 1.
4. $0 \leq h_{ij} \leq 1$ and $\sum_{i=1}^n h_{ii} = p = rank(X)$. Since, the average of the diagonal elements for H is p/n the i^{th} observation is said to be a leverage point if $h_{ii} \geq 2p/n$.
5. Since $\hat{y} = Hy$ we have

$$\hat{y}_i = h_{ii}y_i + \sum_{i \neq j} h_{ij}y_j.$$

   This indicates the importance that y_i has upon $\hat{y}_i$ is given by the magnitude of h_{ii}.

4.2.3 Detection of Influential Observations

If one suspects that the i^{th} observation has an unusual influence upon the prediction equation $\hat{y}$ one can recompute the regression model with the i^{th} observation omitted from the calculation. Suppose that the i^{th} observation is omitted then resultant regression estimate becomes

$$\hat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}y_{(i)}$$

from which we have a new predicted value for y given by $\hat{y}_{(i)} = X\hat{\beta}_{(i)}$.

Cook's Distance

Cook's distance is a measure of how far the original "line" ($\hat{y}$) is from the "new line" ($\hat{y}_{(i)}$) when the i^{th} observation is omitted from the calculation. Cook's distance is

$$\begin{aligned} D_i &= (\hat{y} - \hat{y}_{(i)})'(\hat{y} - \hat{y}_{(i)})/(p\hat{\sigma}^2) \\ &= (\hat{\beta} - \hat{\beta}_{(i)})'X'X(\hat{\beta} - \hat{\beta}_{(i)})/(p\hat{\sigma}^2) \\ &= \left[\frac{\hat{e}_i}{\hat{\sigma}(1-h_{ii})^{1/2}}\right]^2 \left[\frac{h_{ii}}{p(1-h_{ii})}\right]. \end{aligned}$$

DFFITS

A related measure to Cook's distance is the DFFITS statistic given by

$$DFFITS_i^2 = (\hat{\beta} - \hat{\beta}_{(i)})'X'X(\hat{\beta} - \hat{\beta}_{(i)})/(\hat{\sigma}^2_{(i)}).$$

COVRATIO

Another measure of influence is the COVRATIO which is the ratio of the determinant of the covariance matrix for the estimate $\hat{\beta}$, given by $det[\hat{\sigma}^2(X'X)^{-1}]$ when the i^{th} observation has been removed for the computation of the estimate $\hat{\beta}_{(i)}$. That is,

$$COVRATIO = det[\hat{\sigma}^2_{(i)}(X'_{(i)}X_{(i)})^{-1}]/det[\hat{\sigma}^2(X'X)^{-1}].$$

This statistic should be close to one whenever the observation has little influence upon the estimation of β. If the statistics is much different from one then the observation is said to be influential.

Belsley, Kuh, and Welsch suggest that observations with

$$\mid COVRATIO - 1 \mid \geq \frac{3p}{n}$$

where p is the number of parameters in the model and n is the number of observations used to fit the model, are worth investigation.

DFBETAS

The DFBETAS statistics are the scaled measures of the change in each parameter estimate and are calculated by deleting the i^{th} observation:

$$DFBETAS_j = \frac{\hat{\beta}_j - \hat{\beta}_{(i)j}}{\hat{\sigma}_{(i)}\sqrt{(X'X)^{-1}_{jj}}}$$

where $(X'X)_{jj}$ is the $(j,j)^{th}$ element of $(X'X)^{-1}$.

In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter. Belsley, Kuh, and Welsch recommend 2 as a general cutoff value to indicate influential observations and as a size-adjusted cutoff.

Chapter 5

Model Selection

The methods presented in this chapter are used to determine appropriate subset models in the multiple regression problem. Various statistics can be used, including; R^2, the adjusted R^2, s^2 the residual mean square (s^2), and Mallow's C_p statistic.

Mallow's statistics – C_p

The statistics is

$$C_p = RSS_p/s^2 - (n - 2p),$$

where RSS_p is the residual sum of squares from the model containing $p =$ rank(X) parameters (including β_0), and $s^2 = MS_E(r)$ is the residual mean square from the full model containing r predictor variables [assumed to be the most reliable estimate of σ^2]. Note, when $p = r + 1$, $C_p = p$. The idea of using this statistics is to find the smallest value of p such that $C_p \approx p$.

5.1 Subset Selection

Statistical selection procedures include;

Forward Selection

The forward procedure starts with a single variable model (often selected with the highest R^2) and then adds additional variables that satisfy an entry criteria. The process continues until no other variables satisfy the entry criteria.

Backward Selection

The backward procedure starts with the complete (full) model and then eliminates variables that satisfy an exit criteria. The procedure is as follows;

1. Compute the regression model using all the predictor variables.
2. The partial F-value is calculated for every predictor variable using the type II sum of squares.
3. The lowest partial F-value, F_L, is compared with a preselected significance level, F_0.
   (a) If $F_L < F_0$, remove the variable corresponding to F_L, say X_L then recompute the model using the reduced model.

(b) If $F_L > F_0$, adopt the regression model as calculated.

Stepwise Selection

Stepwise is a forward selection method that selects the best single predictor (say, the variable with the largest R^2), X_1 and fit the equation $\hat{y} = f(X_1)$. If this model is not significant, stop and conclude that $\hat{y} = \bar{y}$. If the model is significant, select the next predictor variable, say X_2 based upon the one with the largest partial F-value and the equation is given by $\hat{y} = f(X_1, X_2)$. This model checked for improvement in the R^2 and partial F-values for both variables in the equation. (This differs from the forward procedure in that a first variable may be excluded from the model at this step whereas in forward selection once a variable enters the model it remains). These partial F-values are used to determine whether or not a variable remains in the model or is excluded. The procedure continues until no new variables satisfy the entry criteria.

5.1.1 Ridge Regression

Ridge regression is a popular method for detecting multicollinearity within a regression model. It was first proposed by Hoerl and Kennard (1970) and it was one of the first biased estimation procedures. The idea is fairly simple. Since the matrix $X'X$ is ill-condition or nearly singular one can add positive constants to the diagonal matrix and insure that the resulting matrix is not ill-conditioned. That is, consider the biased normal equations given by

$$(X'X + kI_n)\beta = X'y.$$

With a resulting biased estimate for β given by

$$\tilde{\beta}(k) = (X'X + kI_n)^{-1}X'y,$$

where k is called the shrinkage parameter. Since, $E(\tilde{\beta}) \neq \beta$ some do not want to use such a procedure. However in spite of the fact that $\tilde{\beta}$ is biased, it does have the effect of reducing the variance in the estimator. It can be shown that,

$$var(\hat{\beta}_j) = \sigma^2 1/\lambda_j,$$

where λ_j is the j^{th} eigenvalue of $X'X$. So when $X'X$ is ill-conditioned some of the $\lambda_j's$ are very small, hence $var(\hat{\beta}_j)$ is very large. However,

$$var(\tilde{\beta}_j) = \sigma^2\lambda_j/(\lambda_j + k)^2.$$

Consider the example where $\sigma^2 = 1$, $\lambda_1 = 2.985, \lambda_2 = 0.01$, and $\lambda_3 = 0.005$, the usual least squares estimation gives,

$$\sum_{j=1}^{3} var(\hat{\beta}_j) = \sigma^2 \sum_{i=1}^{3} 1/\lambda_j = .3350 + 100 + 200 = 300.3350.$$

However, if $k = 0.10$ we have,

$$\sum_{j=1}^{3} var(\hat{\beta}_j) = \sigma^2 \sum_{i=1}^{3} \lambda_j/(\lambda_j + k)^2 \approx 2.3.$$

This process of reducing the total variance is very desirable and has led people to proposed similar estimation procedures called shrinkage estimators. In this class, we are interested using this procedure as a way of identifying multicollinearity and the variables which may contribute to this problem.

5.1.2 LASSO

The LASSO procedure is similar to ridge regression in that the estimator is a "shrinkage estimate" whereby one trades off the unbiased property for an estimate that is much more precise (smaller dispersion or mean square error). That is, the least squares estimate is

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\{\|y - X\beta\|^2\}.$$

Whereas the ridge estimate is

$$\hat{\beta}_{ridge} = \underset{\beta}{\text{argmin}}\{\|y - X\beta\|^2 + \lambda \ \|\beta\|^2\}$$

where $\|\beta\| = \sum_{i=1}^{p} \beta_i^2$ is the usual l_2 or Euclidean norm. The LASSO estimate is

$$\hat{\beta}_{lasso} = \underset{\beta}{\text{argmin}}\{\|y - X\beta\|^2 + \lambda \ \|\beta\|\}$$

where $\|\beta\| = \sum_{i=1}^{p} |\beta_i|$ is the l_1 norm.

Hastie, Tibshirani and Friedman book entitled "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" have a very understandable discussion of the above methods. I have reproduced some of their material here.

> In this section we discuss and compare the three approaches discussed so far for restricting the linear regression model: subset selection, ridge regression and the lasso. In the case of an orthonormal input matrix X the three procedures have explicit solutions. Each method applies a simple transformation to the least squares estimate $\hat{\beta}_j$. Ridge regression does a "proportional shrinkage". Lasso translates each coefficient by a constant factor λ, truncating at zero. This is called "soft thresholding," and is used in the context of wavelet-based smoothing in Section 5.9. Best-subset selection drops all variables with coefficients smaller than the Mth largest; this is a form of "hard thresholding."

5.2 Elastic Net Selection

This section applies to the following procedures: REGSELECT.

The METHOD=ELASTICNET option specifies the elastic net method proposed by Zou and Hastie (2005), which bridges the LASSO method and ridge regression. The elastic net method strikes a balance between having a parsimonious model and borrowing strength from correlated regressors, by solving the least squares regression problem with constraints on both the sum of the absolute coefficients and the sum of the squared coefficients.

More specifically, the elastic net coefficients $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$ are the solution to the constrained optimization problem

$$\min\|\mathbf{y} - \mathbf{X}\beta\|^2 \quad \textit{subject to} \quad \sum_{j=1}^{m} \mid \beta_j \mid \ \leq t_1, \sum_{j=1}^{m} \beta_j^2 \leq t_2$$

This can be written as the equivalent Lagrangian form

$$\min\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1 \sum_{j=1}^{m} \mid \beta_j \mid + \lambda_2 \sum_{j=1}^{m} \beta_j^2$$

Elastic net can be treated as a convex combination of LASSO and ridge penalty; pure LASSO and pure ridge are two limiting cases. If t_1 is set to a very large value or, equivalently, if λ_1 is set to 0, then the elastic net

method reduces to ridge regression. If t_2 is set to a very large value or, equivalently, if λ_2 is set to 0, then the elastic net method reduces to LASSO. If t_1 and t_2 are both large or, equivalently, if λ_1 and λ_2 are both set to 0, then the elastic net method reduces to ordinary least squares regression.

The elastic net method can overcome the limitations of LASSO in the following three scenarios:

- If you have more parameters than observations ($m > n$), the LASSO method selects at most n variables before it saturates, because of the nature of the convex optimization problem. This can be a defect for a variable selection method. By contrast, the elastic net method can select more than n variables in this case because of the ridge regression regularization.
- If there is a group of variables that have high pairwise correlations, then whereas LASSO tends to select only one variable from that group, the elastic net method can select more than one variable.
- If you have more observations than parameters ($n > m$), and there are high correlations between predictors, then it has been empirically observed that the prediction performance of LASSO is dominated by ridge regression. In this case, the elastic net method can achieve better prediction performance by using ridge regression regularization.

Part III

Analysis of Variance Models

Chapter 6

One-Way Models with $K > 2$ Populations

In this chapter we leave the regression type models and return to the problem of comparing location parameters for $(k > 2)$ populations.

The first section covers an ANOVA problem that you saw in a stat 2381 course.

6.1 Parametric Models – ANOVA

Consider the simple example given in Kutner Chapter 16, where the objective is to compare the means for 4 designs of retail stores.

Design	Store 1	Store 2	Store 3	Store 4	Store 5
1	11	17	16	14	15
2	12	10	15	19	11
3	23	20	18	17	
4	27	33	22	26	28

The model is

$$\begin{aligned} y_{ij} &= \mu_i + \epsilon_{ij} \\ &= \mu + \tau_i + \epsilon_{ij} \end{aligned}$$

for $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, n_i$ when $a = 4$ and $n_1 = 5, n_2 = 5, n_3 = 4$, and $n_4 = 5$. In this example. a is the number of treatment groups (populations) and n_i is the number of replications for each treatment group. ϵ_{ij} is the unobserved random error term with assumed $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = \sigma^2 > 0$. Note, $\mu_i = \mu + \tau_i$ where μ is the overall mean for y if there were no mean treatment differences and τ_i represents the departure from μ for the i^{th} mean treatment effect.

Let

$$y_{i+} = \sum_{j=1}^{n_i} y_{ij}, \qquad \bar{y}_{i+} = y_{i+}/n_i$$

and

$$y_{++} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}, \qquad \bar{y}_{++} = y_{++}/N,$$

where $N = \sum_{i=1}^{a} n_i$ is the total number of observations in the analysis.

6.1.1 Analysis of the Fixed Effects Model

In this model the objective is to determine if the a treatments effects for the response variable y are equal (on average). That is, one wishes to test the hypothesis,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a$$

versus

$$H_1 : \mu_i \neq \mu_{i'}$$

for at least one pair of treatment means i and i'. In order to use the least squares method when estimating the unknown parameters, μ and τ_i, one needs an additional equation, typically given by $\sum_{i=1}^{a} \tau_i = 0$. Since τ_i represents the i^{th} treatment effect (departure for the overall grand mean given by μ), this equation assumes that the sum of these effects are zero (a perfectly reasonable, albeit unnecessary assumption as we shall see in later chapters). With this assumption the above hypothesis becomes

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$$

versus

$$H_1 : \tau_i \neq 0$$

for some i. Note, in this model the parameters τ_i are assumed to be fixed unknown constants.

Decomposition of the Total Sum of Squares

Assume that $n_i = n$ for $i = 1, 2, \ldots, a$ in the remaining sections for this example. [This assumption is not necessary but it does make the notation simpler]. The total sum of squares is

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} [y_{ij} - \bar{y}_{++} + \bar{y}_{++}]^2 = SS_{CT} + N\bar{y}_{++}^2$$

where $N = a \times n$, $N\bar{y}_{++}^2$ is the correction factor, $\bar{y}_{++} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}$ is the grand or overall mean, and SS_{CT} is the corrected (for μ) total sum of squares given by

$$SS_{CT} = \sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{++})^2.$$

SS_{CT} can be partitioned into two sum of squares, SS_M (model) and SS_E (residual or error)[1],

$$\begin{aligned} SS_{CT} &= \sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{++})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} [(y_{ij} - \bar{y}_{i+}) + (\bar{y}_{i+} - \bar{y}_{++})]^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i+})^2 + n\sum_{i=1}^{a} (\bar{y}_{i+} - \bar{y}_{++})^2 \\ &= SS_E + SS_M. \end{aligned}$$

SS_E can be written as $\sum_{i=1}^{a}(n-1)s_i^2$ where $(n-1)s_i^2 = \sum_{j=1}^{n}(y_{ij} - y_{i+})^2$, the usual sample variance for the i^{th} treatment group. Recall, that if one assumes that the random variable y is normally distributed then the sample variance has a chi-square distribution. If one assumes that the treatment groups are independent of one another then the resultant chi-squares are also independent. Since SS_E is a linear combination of

[1] SS_M is the across treatment sum of squares and SS_E is the average within treatment sum of squares.

chi-squares it shouldn't be surprising to find that its distribution is also chi-square. Before addressing this point consider the expected value of SS_E given by

$$\begin{aligned}
E(SS_E) &= E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i+})^2\right] \\
&= E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}^2-2y_{ij}\bar{y}_{i+}+\bar{y}_{i+}^2)\right] \\
&= E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}y_{ij}^2-\frac{1}{n}\sum_{i=1}^{a}y_{i+}^2\right] \\
&= E\left[\sum_{i=1}^{a}\sum_{j=1}^{n}(\mu+\tau_i+e_{ij})^2-\frac{1}{n}\sum_{i=1}^{a}(\sum_{j=1}^{n}\mu+\tau_i+e_{ij})^2\right] \\
&= N\mu^2+n\sum_{i=1}^{a}\tau_i^2+N\sigma_e^2-N\mu^2-n\sum_{i=1}^{a}\tau_i^2-a\sigma_e^2 \\
&= (N-a)\sigma_e^2
\end{aligned}$$

Note: If one's objective is to find an estimate for σ_e^2 then $\hat{\sigma}_e^2 = MS_E = SS_E/(N-a)$ would be a good candidate.

Using a similar approach (this is for you to do) one can show that

$$E(SS_M) = E\left[n\sum_{i=1}^{a}(\bar{y}_{i+}-\bar{y}_{++})^2\right] = \sigma_e^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{(a-1)}$$

6.1.2 Random Effects Model

This model differs from the fixed effects model by assuming that the treatments are randomly selected from a larger population of many potential treatments where $\tau_i \sim N(0,\sigma_\tau^2)$. The analysis for the random and fixed effects models are the same, however, the interpretations are different. The differences can be seen in the expected value for the treatment sum of squares. That is,

$$E(SS_M) = E\left[n\sum_{i=1}^{a}(\bar{y}_{i+}-\bar{y}_{++})^2\right] = \sigma_e^2 + \frac{n\sigma_\tau^2}{(a-1)}.$$

In the random effects model one is no longer interested in the expected value of the treatments ($\hat{\tau}_i$) but rather in the variance of the treatment effect (σ_τ^2). The hypotheses become

$$H_0 : \sigma_\tau^2 = 0$$

versus

$$H_1 : \sigma_\tau^2 > 0.$$

6.1.3 Statistical Inference

In order to test the null hypothesis one needs to make a distribution assumption concerning the random terms. The assumptions are,

$$e_{ij} \sim N(0,\sigma_e^2)$$

which implies

$$y_{ij} \sim N(\mu + \tau_i, \sigma_e^2) \quad \text{fixed effects model}$$

and

$$y_{ij} \sim N(\mu, \sigma_\tau^2 + \sigma_e^2) \quad \text{random effects model.}$$

Cochran's Theorem can be used to test the null hypothesis.

Theorem 6.1 (Cochran's Theorem) *Let $Z_i \sim N(0,1)$ for $i = 1, 2, \ldots, v$ and*

$$\sum_{i=1}^{v} Z_i^2 = \sum_{i=1}^{s} Q_i$$

where $s \leq v$ and $E[Q_i] = v_i$. Then $Q_1, \ldots, Q_s$ are independent and $Q_i \sim \chi^2(df = v_i)$ if and only if $v = \sum_{i=1}^{s} v_i$.

By noting that

$$\frac{SS_{CT}}{\sigma_e^2} = \frac{SS_M}{\sigma_e^2} + \frac{SS_E}{\sigma_e^2}$$

and

$$\frac{SS_{CT}}{\sigma_e^2} = \sum_{i=1}^{a}\sum_{j=1}^{n} \frac{(y_{ij} - \bar{y}_{++})^2}{\sigma_e^2} = \sum_{i=1}^{a}\sum_{j=1}^{n} (\frac{y_{ij}}{\sigma_e})^2 - \left(\frac{\bar{y}_{++}}{\frac{\sigma_e}{\sqrt{N}}}\right)^2$$

we see that

$$\frac{SS_{CT}}{\sigma_e^2} \sim \chi^2(df = N - 1).$$

From Cochran's Theorem we have,

$$\frac{SS_E}{\sigma_e^2} \sim \chi^2(df = N - a)$$

and

$$\frac{SS_M}{\sigma_e^2} \sim \chi^2(df = a - 1).$$

In which case, we have

$$F = \frac{MS_M}{MS_E} \sim F(ndf = a - 1, ddf = N - a)$$

where

$$MS_{tr} = SS_M/(a - 1)$$

and

$$MS_E = SS_E/(N - a).$$

The above holds whenever SS_E and SS_M are independent. The independence follows from the random sample assumption made on the error structure. The independence of SS_E and SS_M does not hold when the errors are dependent.

Note that

$$E[F] = 1 + \frac{n\sum_{i=1}^{a} \tau_i^2}{(a-1)\sigma_e^2} \quad \text{fixed effects model}$$

and

$$E[F] = 1 + \frac{n\sigma_\tau^2}{(a-1)\sigma_e^2} \quad \text{random effects model.}$$

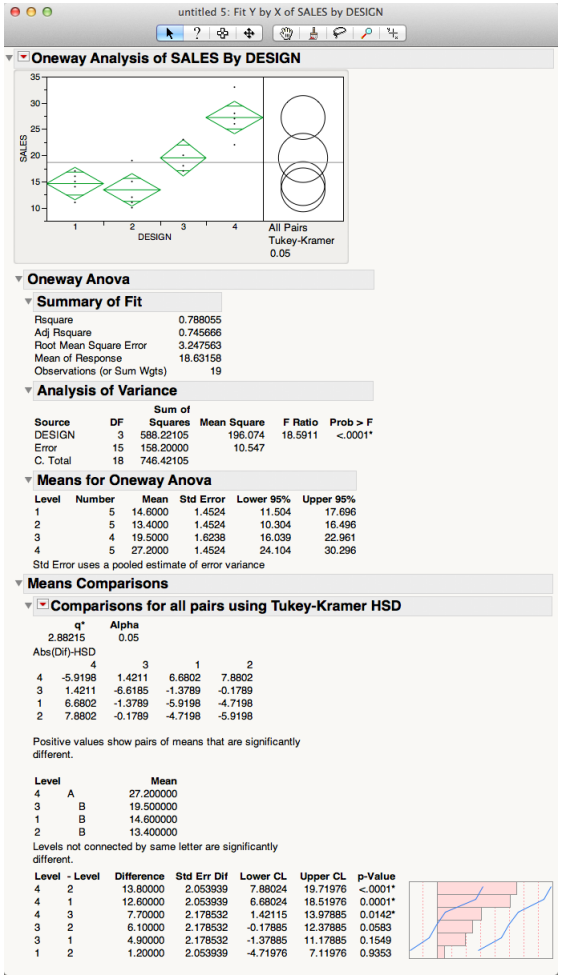
In either case, $E[F] \approx 1$ when the null hypothesis is true. Therefore, one rejects the null hypothesis in either model when $F >> 1$.

The results are summarized in the ANOVA table

Source of Variation	Sum of Squares	Degrees of Feedom	Mean Square	F
Treatment	SS_M	a - 1	MS_M	$F = \frac{MS_M}{MS_E}$
Error	SS_E	N - a	MS_E	
Corrected Total	SS_{CT}	N - 1		

Example

The results using JMP for the Kutner example in Chapter 16.1 is given by,



6.2 Multiple Comparisons

A statistically significant ANOVA F-test indicates differences in the mean effects, but it does not indicate where the differences are found. Multiple-comparison procedures (MCP) provide detailed information about

the observed differences among the means. The goal in multiple comparisons is to compare the average effects of three or more "treatments" to decide which treatments are better, which ones are worse, and by how much, while controlling the probability of making an incorrect decision (overall α-level).

Hsu (1996) categorized MCP two ways:

1. By the type comparisons made:
   - comparisons between all pairs of means
   - comparisons between a control and all other means
2. By the strength of inference provided. The strength of inference is related to what type of error rate the MCP controls. The types of inference (from weakest to strongest) are:
   - Individual – differences between means, unadjusted for multiplicity
   - Inhomogeneity – means are different
   - Inequalities – which means are different
   - Intervals – simultaneous confidence intervals for mean differences

Methods that control only individual error rates are not true MCP. Methods that yield the strongest level of inference, simultaneous confidence intervals, are preferred, since they enable you not only to say which means are different but also to put confidence bounds on *how much* they differ, making it easier to assess the practical significance of a difference.

6.2.1 Pairwise Comparisons

The methods discussed in this section depend on the standardized pairwise differences $t_{ij} = (\bar{y}_i - \bar{y}_j)/\hat{\sigma}_{ij}$, where

- i and j are the indices of two groups
- n_i and n_j are the sample sizes of two groups
- $\bar{y}_i$ and $\bar{y}_j$ are the means or LS-means for groups i and j
- $\hat{\sigma}_{ij}$ is the estimated standard error of $\bar{y}_i - \bar{y}_j$.
  - MEANS – simple arithmetic means, $\hat{\sigma}^2_{ij} = s^2(1/n_i + 1/n_j)$, where s^2 is the mean square for error with ν degrees of freedom.
  - LSMEANS – the linear combinations $\mathrm{l}'_i\mathbf{b}$ and $\mathbf{l}'_j\mathbf{b}$ of the parameter estimates, $\hat{\sigma}^2_{ij} = s^2\mathbf{l}'_i(\mathbf{X}'\mathbf{X})^-\mathbf{l}_j$.

The significance tests for the methods are of the form

$$|t_{ij}| \geq c(\alpha)$$

where $c(\alpha)$ is a constant depending on the significance level. The corresponding confidence intervals are of the form

$$[(\bar{y}_i - \bar{y}_j) \pm \hat{\sigma}_{ij}c(\alpha)]$$

The simplest approach to multiple comparisons is to do a t-test on every pair of means. For the i^{th} and j^{th} means, reject the null hypothesis that the population means are equal if

$$|t_{ij}| \geq t(\alpha; \nu)$$

where α is the significance level, ν is the number of error degrees of freedom, and $t(\alpha;\nu)$ is the two-tailed critical value from a Student's t-distribution. If the cell sizes are all equal to, say, n, the preceding formula can be rearranged to give

$$|\bar{y}_i - \bar{y}_j| \geq t(\alpha;\nu)s\sqrt{\frac{2}{n}}.$$

This statistics is called **Fisher's least significant difference (LSD)**.

There is a problem with repeated t-tests. Suppose there are 10 means and each t-test is performed at the 0.05 level. There are $10(10-1)/2 = 45$ pairs of means to compare, each with a 0.05 probability of a type 1 error (a false rejection of the null hypothesis). The chance of making at least one type 1 error is much higher than 0.05. It is difficult to calculate the exact probability, but you can derive a pessimistic approximation by assuming that the comparisons are independent, giving an upper bound to the probability of making at least one type 1 error (the experiment-wise error rate) of

$$1 - (1 - 0.05)^{45} = 0.90.$$

The actual probability is less than 0.90. As the number of means increases, the chance of making at least one type 1 error approaches 1.

If you decide to control the individual type 1 error rates for each comparison, you are controlling the individual or comparison-wise error rate. If you want to control the overall type 1 error rate for all the comparisons, you are controlling the experiment-wise error rate. Statistical methods for comparing three or more means while controlling the probability of making at least one type 1 error are called *multiple-comparison procedures* (MCP).

It has been suggested that the experiment-wise error rate can be held to the α level by performing the overall ANOVA F test at the α level and making further comparisons only if the F test is significant, as in Fisher's protected LSD. This assertion is false if there are more than three means (Einot and Gabriel, 1975). The following abbreviations are used in the discussion:

- CER – comparison-wise error rate
- EERC – experiment-wise error rate under the complete null hypothesis
- MEER – maximum experiment-wise error rate under any complete or partial null hypothesis

These error rates are associated with the different *strengths of inference*:

- Individual tests control the CER.
- Tests for inhomogeneity of means control the EERC.
- Tests that yield confidence inequalities or confidence intervals control the MEER.

Note: A preliminary F test controls the EERC but not the MEER.

You can control the MEER at the α level by setting the CER to a sufficiently small value.

Bonferroni Method

The Bonferroni inequality (Miller, 1981) can be used for this purpose. If CER $= \alpha/c$ then the MEER is less than α where c is the total number of comparisons. The Bonferroni t test with MEER $< \alpha$ declares two means to be significantly different if

$$|t_{ij}| \geq t(\epsilon;\nu)$$

where

$$\epsilon = \frac{\alpha}{\binom{k}{2}} = \frac{2\alpha}{k(k-1)}$$

for the comparison of k means.

Sidak Method

Sidak (1967) has provided a tighter bound. If CER $= 1 - (1 - \alpha)^{1/c}$ then MEER $\leq \alpha$ where c is the total number of comparisons. The Sidak t test (Games, 1977) is

$$|t_{ij}| \geq t(\epsilon; \nu)$$

where

$$\epsilon = 1 - (1 - \alpha)^{\frac{2}{k(k-1)}}$$

for the comparison of k means.

Scheffe Method

Scheffe (1953, 1959) proposes a method to control the MEER for any set of contrasts. Two means are declared significantly different if

$$|t_{ij}| \geq \sqrt{df \cdot F(\alpha; df, \nu)}$$

where $F(\alpha; df, \nu)$ is the α-level critical value of an F distribution with df numerator degrees of freedom and ν denominator degrees of freedom. The value of df is $k - 1$ for the MEANS statement. For the LSMEANS statement, df is the rank of the contrast matrix L for LS-means differences. In more general contexts df is the rank of the contrast covariance matrix $LCov(b)L'$.

Scheffe's test is compatible with the overall ANOVA F test in that Scheffe's method never declares a contrast significant if the overall F test is nonsignificant. Most other multiple-comparison methods can find significant contrasts when the overall F test is nonsignificant and, therefore, suffer a loss of power when used with a preliminary F test. Scheffe's method might be more powerful than the Bonferroni or Sidak method if the number of comparisons is large relative to the number of means. For pairwise comparisons, Sidak t tests are generally more powerful.

Tukey Method

Tukey (1952, 1953) proposes a test designed specifically for pairwise comparisons based on the studentized range. This procedure controls the MEER when the sample sizes are equal. Tukey (1953) and Kramer (1956) independently propose a modification for unequal cell sizes. This method has fared extremely well in Monte Carlo studies (Dunnett, 1980). In addition, Hayter (1984) gives a proof that the Tukey-Kramer procedure controls the MEER for means comparisons, and Hayter (1989) describes the extent to which the Tukey-Kramer procedure has been proven to control the MEER for LS-means comparisons. The Tukey-Kramer method is more powerful than the Bonferroni, Sidak, or Scheffe's method for pairwise comparisons. Two means are considered significantly different by the Tukey-Kramer criterion if

$$|t_{ij}| \geq q(\alpha; k, \nu)$$

where $q(\alpha; k, \nu)$ is the α-level critical value of a studentized range distribution of k independent normal random variables with ν degrees of freedom.

Hochberg Method

Hochberg (1974) devised a method the is similar to Tukey's, but it uses the studentized maximum modulus instead of the studentized range and employs the uncorrelated t inequality of Sidak (1967). It has been shown to hold the MEER at a level not exceeding α with unequal sample sizes. It is generally less powerful than the Tukey-Kramer method and always less powerful than Tukey's test for equal cell sizes. Two means are declared significantly different if

$$|t_{ij}| \geq m(\alpha; c, \nu)$$

where $m(\alpha; c, \nu)$ is the α-level critical value of the studentized maximum modulus distribution of c independent normal random variables with ν degrees of freedom and $c = k(k-1)/2$.

Gabriel Method

Gabriel (1978) proposes another method based on the studentized maximum modulus. This method is applicable only to arithmetic means. Reject H_0 if

$$\frac{|\bar{y}_i - \bar{y}_j|}{s\left(\frac{1}{\sqrt{2n_i}} + \frac{1}{\sqrt{2n_j}}\right)} \geq m(\alpha; k, \nu)$$

For equal cell sizes, Gabriel's test is equivalent to Hochberg's GT2 method. For unequal cell sizes, Gabriel's method is more powerful than GT2 but might become liberal with highly disparate cell sizes (see also Dunnett (1980)). Gabriel's test is the only method for unequal sample sizes that lends itself to a graphical representation as intervals around the means. Assuming $\bar{y}_i > \bar{y}_j$, you can rewrite the preceding inequality as

$$\bar{y}_i - m(\alpha; k, \nu)\frac{s}{\sqrt{2n_i}} \geq \bar{y}_j + m(\alpha; k, \nu)\frac{s}{\sqrt{2n_j}}$$

The expression on the left does not depend on j, nor does the expression on the right depend on i. Hence, you can form what Gabriel calls an (l, u)-interval around each sample mean and declare two means to be significantly different if their (l, u)-intervals do not overlap.

6.2.2 Comparing All Treatments to a Control

A special case of means comparison is when all treatments are compared to a single control. In this case, you can achieve better power by using a method that is restricted to test only

$$|t_{i0}| \geq d(\alpha; k, \nu, \rho_1, \ldots, \rho_{k-1})$$

where $\bar{y}_0$ is the control mean and $d(\alpha; k, \nu, \rho_1, \ldots, \rho_{k-1})$ is the critical value of the "many-to-one t-statistics" (Miller; 1981; Krishnaiah and Armitage; 1966) for k means to be compared to a control, with ν error degrees of freedom and correlations $\rho_1, \ldots, \rho_{k-1}$, $\rho_i = n_i/(n_0 + n_i)$. The correlation terms arise because each of the treatment means is being compared to the same control. **Dunnett's** test holds the MEER to a level not exceeding the stated α.

6.3 Nonparametric Methods

6.3.1 Kruskal-Wallis Test

Assume that one has a random samples of size n_i from each of $k > 2$ populations, where X_{ij} is the j^{th} sample (observation) from population i. Assume that X_{ij} are at least ordinal data from a distribution with $F_X(x \mid \theta_i)$ where the median for the i^{th} class is $\theta_i = \theta + \tau_i$ and θ is the grand median. The test of interest is that

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_k = 0$$

versus the alternative that there exists at least one inequality. The Kruskal-Wallis[2] procedure is as follows:

1. Rank the combined data of $N = \sum_{i=1}^{k} n_i$ observations where $R_{ij} = rank(X_{ij}) \in \{1, 2, \ldots, N\}$.

[2]SAS-PROC NPAR1WAY

2. Compute

$$\begin{aligned} S &= \frac{12}{N(N+1)} \sum_{i=1}^{k} n_j (\bar{R}_{i+} - R_{++})^2 \\ &= \left(\frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_{i+}^2}{n_i} \right) - 3(N+1) \end{aligned}$$

where $R_{i+} = \sum_{j=1}^{n_i} R_{ij}$, $\bar{R}_{i+} = \frac{R_{i+}}{n_i}$, $R_{++} = \frac{N+1}{2}$.

3. Reject H_0 if S is too large where the asymptotic distribution is chi-square with degrees of freedom $df = k - 1$.

There are procedures for handling ties and multiple comparisons whenever the null hypothesis is rejected.

6.3.2 Jonckheere-Terpstra Test

The Jonckheere-Terpstra test[3] is a nonparametric test for ordered differences among classes. It tests the null hypothesis that the distribution of the response variable does not differ among classes. It is designed to detect alternatives of ordered class differences, which can be expressed as $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$ (or $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_k$), with at least one of the inequalities being strict, where τ_i denotes the effect of class i and k is the number of group(rows), R, in the table.[4]

The Jonckheere-Terpstra test statistic is computed by forming $\binom{k}{2} = k(k-1)/2$ Mann-Whitney counts $M_{i,i'}$, where $i < i'$, for pairs of rows in the contingency table,

$$\begin{aligned} M_{i,i'} &= \{\text{number of times } X_{ij} < X_{i'j'}, \quad j = 1, \ldots, n_{i+}; \quad j' = 1, \ldots, n_{i'+} \} \\ &\quad + \frac{1}{2} \{\text{number of times } X_{ij} = X_{i'j'}, \quad j = 1, \ldots, n_{i+}; \quad j' = 1, \ldots, n_{i'+} \} \end{aligned}$$

where X_{ij} is response j in row i. The Jonckheere-Terpstra test statistic is computed as

$$J = \sum_{1 \leq i < i' \leq k} \sum M_{i,i'}.$$

This test rejects the null hypothesis of no difference among classes for large values of J. Asymptotic p-values for the Jonckheere-Terpstra test are obtained by using the normal approximation for the distribution of the standardized test statistic. The standardized test statistic is computed as

$$J^* = (J - \mathrm{E}_0(J)) \ / \ \sqrt{\mathrm{Var}_0(J)}$$

where $\mathrm{E}_0(J)$ and $\mathrm{Var}_0(J)$ are the expected value and variance of the test statistic under the null hypothesis,

$$\mathrm{E}_0(J) = \left(n^2 - \sum_i n_{i+}^2 \right) / 4$$

[3] PROC FREQ with the JT option in the TABLES statement

[4] For such ordered alternatives, the Jonckheere-Terpstra test is more powerful than a test designed for general class difference alternatives, such as the Kruskal-Wallis test. See Pirie (1983) and Hollander and Wolfe (1999) for more information about the Jonckheere-Terpstra test.

$$\text{Var}_0(J) = A/72 + B/\left(36n(n-1)(n-2)\right) + C/\left(8n(n-1)\right)$$

where

$$A = n(n-1)(2n+5) - \sum_i n_{i+}(n_{i+}-1)(2n_{i+}+5) - \sum_j n_{+j}(n_{+j}-1)(2n_{+j}+5)$$

$$B = \left(\sum_i n_{i+}(n_{i+}-1)(n_{i+}-2)\right)\left(\sum_j n_{+j}(n_{+j}-1)(n_{+j}-2)\right)$$

$$C = \left(\sum_i n_{i+}(n_{i+}-1)\right)\left(\sum_j n_{+j}(n_{+j}-1)\right)$$

In the above procedure one must compute $\binom{k}{2} = k(k-1)/2$ Mann-Whitney statistics. Randles and Wolfe (1991) present a method for computing the Jonckheere-Terpstra statistics using $k-1$ Mann-Whitney statistics.

6.3.3 Randles Modification of the JT Test

Let U_s for $s = 2, \ldots, k$ denote the Mann-Whitney test statistic between the first $(s-1)$ samples combined and the s^{th} sample. Then

$$J = \sum_{s=2}^{k} U_s$$

where $U_2, U_3, \ldots, U_k$ are independent when the null hypothesis is true. [Odeh (1971)].[5]

6.3.4 Friedman's Test

Suppose that one has a block effect in the Kruskal-Wallis model. That is, let $E(X_{ij}) = \mu + \tau_i + \beta_j$ for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$. Let $N = nk$. The test of interest is that

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_k$$

versus the alternative that there exists at least one inequality. Friedman's procedure[6] is as follows:

1. Rank the data within each block where $R_{ij} = rank(X_{ij}) \in \{1, 2, \ldots, k\}$.
2. Compute

$$\begin{aligned} S &= \frac{12n}{k(k+1)} \sum_{i=1}^{k} (\bar{R}_{i+} - R_{++})^2 \\ &= \left(\frac{12}{nk(k+1)} \sum_{i=1}^{k} R_{i+}^2\right) - 3n(k+1) \end{aligned}$$

where $R_{i+} = \sum_{j=1}^{n} R_{ij}$, $\bar{R}_{i+} = \frac{R_{i+}}{n}$, $R_{++} = \frac{k+1}{2}$.

[5] The independence of $U_2, U_3, \ldots, U_k$ enables one to easily compute the mean and variance of J.
[6] PROC FREQ with score = rank.

3. Reject H_0 if S is too large where the asymptotic distribution is chi-square with degrees of freedom $df = k - 1$.

There are procedures for handling ties and multiple comparisons whenever the null hypothesis is rejected. Furthermore, there is a test similar to the Joncheere-Terpstra (JT) procedure with the blocked design.

Chapter 7

ANOVA Models – Matrix Notation

Consider the one-way ANOVA model

$$y_{ij} = \mu + \tau_i x_{ij} + \epsilon_{ij} \tag{7.1}$$

for $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, n_i$, and $N = \sum_{i=1}^{a} n_i$ where ϵ_{ij} is the unobserved distance that the observation y_{ij} is from the population mean for the i^{th} treatment group, $E[y_{ij}] = \mu_i = \mu + \tau_i$. The usual linear model is the ANOVA model when the independent variable X_{ij} is an indicator variable, i.e., $x_{ij} = 1$ if the observation is from the i^{th} treatment group and is 0, otherwise. The model can be written as

$$\mathbf{y} = X\beta + \vec{\epsilon} \tag{7.2}$$

where

$$\tilde{\mathbf{y}} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{an_a} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{j}_{n_1} & \mathbf{j}_{n_1} & 0 & \cdots & 0 \\ \mathbf{j}_{n_2} & 0 & \mathbf{j}_{n_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{j}_{n_a} & 0 & 0 & \cdots & \mathbf{j}_{n_a} \end{pmatrix} \qquad \vec{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{an_a} \end{pmatrix}$$

and

$$\vec{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_a \end{pmatrix} \qquad \mathbf{j}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

7.1 Least Squares

As before, the least squares problem consists of determining $\hat{\beta}$ that minimizes

$$\begin{aligned} Q(\beta) &= (y - X\beta)'(y - X\beta) \\ &= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \\ &= y'y - 2\beta'(X'y) + \beta'(X'X)\beta \end{aligned}$$

The score equation becomes

$$\frac{\partial Q(\beta)}{\partial \beta} = -2X'y + 2X'X\beta = 0$$

with the normal equations given by

$$X'X\beta = X'y. \tag{7.3}$$

In the one-way ANOVA example X is $N \times (a+1)$ with $\text{rank}(X) = a$, in which case $(X'X)^{-1}$ does not exist. Hence, there is not an unique solution to equation (7.3). There have been several approaches to finding an unique solution to this problem. One method is to add another equation, such as, assume that $\sum_{i=1}^{a} \tau_i = 0$. A second approach is to use the *cell means model* of using μ_i rather than $\mu + \tau_i$. A third approach and the one predominately used in modern statistical software[1] is to define a non-unique solution given by

$$\tilde{\beta} = (X'X)^{-}X'y \tag{7.4}$$

where $(X'X)^{-}$ is a generalized inverse of $(X'X)$ satisfying

$$(X'X)(X'X)^{-}(X'X) = (X'X).$$

Properties of the generalized inverse (G-inverse) are given at the end of these notes.

7.1.1 Properties of the G-inverse Solution

Since $(X'X)^{-}$ is not unique, it follows that $\tilde{\beta}$ is not unique. However, we can derive the following results since $\tilde{\beta}$ is a linear function of the normal random variable y:

- Expected value of $\tilde{\beta}$
$$E[\tilde{\beta}] = (X'X)^{-}X'E[y] = (X'X)^{-}X'X\beta \neq \beta.$$
- Variance of $\tilde{\beta}$

  $Var(\tilde{\beta}) = Var((X'X)^{-}X'y) = (X'X)^{-}X'Var(y)X(X'X)^{-'} = (X'X)^{-}X'X(X'X)^{-'}\sigma^2 \neq \sigma^2(X'X)^{-1}.$
- Estimate for $E[y]$
$$\widetilde{E[y]} = X(X'X)^{-}X'y = Hy$$
  where $H = X(X'X)^{-}X'$ the *Hat matrix*[2]
- SS_E
$$SS_E = (y - X\tilde{\beta})'(y - X\tilde{\beta}) = y'(I-H)y = y'y - \tilde{\beta}'X'y.$$
- $E[SS_E]$
$$\begin{aligned} E[SS_E] = E[y'(I-H)y] &= tr[(I-H)I\sigma^2] + \beta'X'(I-H)X\beta \\ &= \sigma^2 tr[(I-H)] = \sigma^2(N - rank(X)) \\ &= (N-a)\sigma^2. \end{aligned}$$

  It follows that

  1. $(I-H)$ and H are idempotent matrices and it follows that $H(I-H) = (I-H)H = 0$.
  2. $X'(I-H)X = X'X - X'HX = X'X - X'\Big[X(X'X)^{-}X'X\Big] = X'X - X'\Big[X\Big] = 0$, since $(X'X)^{-}X'$ is a generalized inverse of X for any choice of $(X'X)^{-}$.[3]

  In which case, one has an unbiased estimate of σ^2 as

$$\hat{\sigma}^2 = SS_E/(N - rank(X)) = SS_E/(N-a).$$

[1] SAS PROC GLM.

[2] Due to the fact that one has the generalized inverse of a symmetric matrix $(X'X)$ it turns out the H is invariant to the choice of $(X'X)^{-}$ which means that $\widetilde{E[y]}$ is the same for all choices of the generalized inverse of $(X'X)$.

[3] I haven't proved this, so a proof would be instructive.

- Partitioning of the total sum of squares

1	2	3
	SS_m	
SS_{tr}	$SS_{tr\|m}$	$SS_{tr\|m}$
SS_E	SS_E	SS_E
SS_{total}	SS_{total}	$SS_{c.total}$

where

$$SS_m = N\bar{y}_{\cdot\cdot}^2 = \frac{1}{N} y' j_N j_N' y$$

$$SS_{tr} = y'Hy \qquad SS_{tr|m} = y'(H - \frac{1}{N} j_N j_N')y = SS_{tr} - SS_m$$

and

$$SS_E = y'(I-H)y \quad SS_{total} = y'y \quad SS_{c.total} = SS_{total} - SS_m$$

- Coefficient of determination

$$R^2 = \frac{SS_{tr} - SS_m}{SS_{total} - SS_m} = \frac{SS_{tr|m}}{SS_{c.total}}.$$

7.2 Distributional Properties

As before it is necessary to make distributional assumptions in order to make any inference about the population parameters. Assume that the unobserved error $\epsilon_i, i = 1, 2, \ldots, N$ are i.i.d normals with mean = 0 and variance = σ^2 or

$$\vec{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{an_a} \end{pmatrix} \sim Gaussian_N(0, \sigma^2 I_N).$$

Using the properties of linear transformations of normal variables, one has:

- y is normal

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{an_a} \end{pmatrix} \sim Gaussian_N(X\beta, \sigma^2 I_N).$$

- $\tilde{\beta}$ is normal

$$\hat{\beta} \sim N_{a+1}((X'X)^-(X'X)\beta, \sigma^2(X'X)^- X'X(X'X)^-).$$

- $\tilde{\beta}$ and $\hat{\sigma}^2$ are independent

$$\tilde{\beta} = (X'X)^- X'y = By$$

and

$$SS_E = y'(I-H)y = y'Ay.$$

SS_E and $\tilde{\beta}$ are independent since

$$BA = (X'X)^- X'(I-H) = (X'X)^-(X'-X'H) = (X'X)^-(X'-X'X(X'X)^- X') = (X'X)^-(X'-X') = 0$$

and $X(X'X)^-$ is a generalized inverse for X'.

- $SS_E/\sigma^2 \sim \chi^2(df = N - rank(X))$

$$SS_E/\sigma^2 = y'(I-H)y/\sigma^2 = z'(I-H)z$$

where $z = y/\sigma \sim N_N(\mu_z = X\beta/\sigma, I_n)$. Using the distributional properties of quadratic forms we have

$$SS_E/\sigma^2 \sim \chi^2(df = rank(I-H), \lambda = \beta' X'(I-H)X\beta/2\sigma^2) \sim \chi^2(N - rank(X) = N - a).$$

- Distributions of SS_{tr}, SS_M and $SS_{tr|M}$. Using the properties of the distributions of quadratic forms it follows that

$$SS_{tr}/\sigma^2 = y'Hy/\sigma^2 \sim \chi^2(df = rank(H) = rank(X), \lambda = \beta X'X\beta/2\sigma^2),$$

$$SS_M/\sigma^2 = \frac{1}{\sigma^2}N\bar{y}_{\cdot\cdot}^2 = \frac{1}{N\sigma^2}y'j_N j_N'y \sim \chi^2(df = 1, \lambda = \frac{1}{2N\sigma^2}(j_N'X\beta)^2),$$

and

$$SS_{tr|m}/\sigma^2 = \frac{1}{\sigma^2}y'(H - \frac{1}{N}j_N j_N')y \sim \chi^2(df = rank(X) - 1, \lambda = \frac{1}{2\sigma^2}\beta' X'(I - \frac{1}{N}j_N j_N')X\beta).$$

7.2.1 One Way ANOVA Table

As in the previous chapter one has the analysis of variance table for the simple one-way model is given by

Source	Sum of Squares	Degrees of Freedom	Mean Square
model	$SS_{tr} = y'Hy$	rank(X) = a	$MS_{tr} = SS_{tr}/a$
Residual	$SS_E = y'(I-H)y$	N - rank(X) = N - a	$MS_E = SS_E/(N-a)$
Total	$y'y$	N	

The model is not usually summarized in this form but is usually given as

Source	Sum of Squares	Degrees of Freedom	Mean Square
mean	$SS_m = \frac{1}{N}y'j_N j_N'y$	1	
adjusted model	$SS_{tr\|m} = SS_{tr} - SS_m$	rank(X) - 1 = a -1	$MS_{tr\|m} = SS_{tr\|m}/(a-1)$
Residual	$SS_E = y'(I-H)y$	N - rank(X) = N - a	$MS_E = SS_E/(N-a)$
Total	$y'y$	N	

or more commonly as

Source	Sum of Squares	Degrees of Freedom	Mean Square
adjusted model	$SS_{tr\|m} = SS_{tr} - SS_m$	rank(X) - 1 = a - 1	$MS_{tr\|m} = SS_{tr\|m}/(a-1)$
Residual	$SS_E = y'(I-H)y$	N - rank(X) = N - a	$MS_E = SS_E/(N-a)$
Corrected Total	$y'y - SS_m$	N - 1	

7.2.2 Estimable Functions

Since $\tilde{\beta}$ is not unique, how and why is it useful? The answer to this question depends on the answer to the question, are there linear combinations of β that can be uniquely estimated with a linear combination of $\tilde{\beta}$?

Definition: 7.1 (Estimable Functions) *A linear function of the parameters is said to be estimable if it is equal to some linear function of the expected value of y. That is, $l = q'\beta$ is estimable if there exists a vector t such that $l = q'\beta = t'E[y]$. Note: t need not be unique.*

The properties of estimable functions are;

- If $l = q'\beta$ is estimable, then $q'\beta = t'E[y] = t'X\beta$ for some t. Since this holds for all β it follows that
$$q' = t'X$$
for some t. This means that $q \in \mathcal{C}(X)$ the column space of the matrix X.
- A given function $l = q'\beta$ is estimable if $q'(X'X)^-X'X = q'$. This follows since there exists a vector t satisfying $q' = t'X$ hence $q'(X'X)^-X'X = t'X(X'X)^-X'X = t'X = q'$.
- The expected value of any observation is estimable. Hence, functions like $E[y_{1j}] = \mu + \tau_1$ and $E[y_{2j}] = \mu + \tau_2$ are estimable. However, τ_i is not estimable.
- Any linear combination of estimable functions is estimable. Hence, $(\mu + \tau_1) - (\mu + \tau_2) = \tau_1 - \tau_2$ is estimable.
- When $l = q'\beta$ is estimable, then l is invariant to the choice of $\tilde{\beta}$, since
$$q'\beta = t'X\tilde{\beta} = t'X(X'X)^-X'y = t'Hy$$
and H is invariant to the choice of $(X'X)^-$.
- The best linear unbiased estimate (b.l.u.e.) of $l = q'\beta$ is $q'\tilde{\beta}$.
- Confidence intervals for $l = q'\beta$ is
$$q'\tilde{\beta} \pm \hat{\sigma} t_{\alpha/2}(N - rank(X))\sqrt{q'(X'X)^-q}.$$

7.2.3 Testable Functions

The hypothesis $H_0 : K'\beta = m$ is testable if the function $K'\beta$ is estimable. Using the properties of linear transformation of normal variates, $y \sim N(X\beta, \sigma^2 I_N)$, enables one to have

- $K'\tilde{\beta} - m \sim N(K'\beta - m, K'(X'X)^-K\sigma^2)$
- $Q/\sigma^2 \sim \chi^2[df = s, \lambda = (K'\beta - m)'(K'(X'X)^-K)^{-1}(K'\beta - m)/2\sigma^2]$, where $s = rank(K'(X'X)^-K)$ and
$$Q = (K'\tilde{\beta} - m)'(K'(X'X)^-K)^{-1}(K'\tilde{\beta} - m).$$

From which one has the following table when $m = 0$

Source	Sum of Squares	Degrees of Freedom
Full adjusted model	$SS_{tr\|m} = SS_{tr} - SS_m$	rank(X) - 1 = a - 1
$H : K'\beta = 0$	Q	s
Reduced restricted model	$SS_{tr\|m} - Q$	(a - s - 1)
Residual	SS_E	N - a
Total	$y'y - SS_m$	N - 1

The tests of hypothesis are;

- Test statistic for the full model,
$$\frac{SS_{tr|m}/(a-1)}{SS_E/(N-a)}$$
- Test statistic for the hypothesis $K'\beta = 0$,
$$\frac{Q/s}{SS_E/(N-a)}$$
- Test statistic for the full model under the assumption $K'\beta = 0$,
$$\frac{(SS_{tr|m} - Q)/(a-s-1)}{SS_E/(N-a)}$$

7.2.4 Independent and Orthogonal Contrasts

The numerator for the hypothesis $H_0 : K'\beta = 0$ is

$$Q = \tilde{\beta}'K(K'(X'X)^{-}K)^{-1}K'\tilde{\beta}.$$

Suppose that the rows of K are k_i and one has testable hypothesis $H_0 : k_i'\beta = 0$, for $i = 1, 2, \ldots, s = rank(K)$ and the numerators for these hypotheses are

$$q_i = \tilde{\beta}'k_i(k_i'(X'X)^{-}k_i)^{-1}k_i'\tilde{\beta}$$

for $i = 1, \ldots, s$. Under what conditions are q_i and q_j independent and $Q = \sum_{i=1}^{s} q_i$? Note: if these conditions hold then the estimable functions given by $k_i'\beta$ are said to be linearly independent orthogonal contrasts. A sufficient condition is

$$k_i'(X'X)^{-}(X'X)(X'X)^{-}k_j = 0.$$

Since $k_i'(X'X)^{-}X'X = k_i'$, it follows that one has linearly independent orthogonal contrasts whenever

$$k_i'(X'X)^{-}k_j = 0.$$

Note: in general two vectors, k_i and k_j, are orthogonal if $k_i'k_j = 0$.

7.2.5 SAS Four Type of Estimable Functions

SAS has four types of estimable functions, denoted as type I - IV.

Type I

Type I sums of squares (SS), also called **sequential sums of squares**, are the incremental improvement in error sums of squares as each effect is added to the model. They can be computed by fitting the model in steps and recording the difference in error sum of squares at each step.

Source	Type I SS
A	$SS(A \mid \mu)$
B	$SS(B \mid \mu, A)$
A*B	$SS(A * B \mid \mu, A, B)$

Type I sums of squares are displayed by default because they are easy to obtain and can be used in various hand calculations to produce sum of squares values for a series of different models. Nelder (1994) and others have argued that Type I and II sums are essentially the only appropriate ones for testing ANOVA effects; however, refer also to the discussion of Nelder's article, especially Rodriguez et al. (1995) and Searle (1995).

The Type I hypotheses have these properties:

- Type I sum of squares for all effects add up to the model sum of squares. None of the other sum of squares types have this property, except in special cases.
- Type I hypotheses can be derived from rows of the Forward-Dolittle transformation of X'X (a transformation that reduces X'X to an upper triangular matrix by row operations).
- Type I sum of squares are statistically independent of each other under the usual assumption that the true residual errors are independent and identically normally distributed.
- Type I hypotheses depend on the order in which effects are specified in the MODEL statement.
- Type I hypotheses are uncontaminated by parameters corresponding to effects that precede the effect being tested; however, the hypotheses usually involve parameters for effects following the tested effect in the model. For example, in the model

  $$Y = A\ B;$$

  the Type I hypothesis for B does not involve A parameters, but the Type I hypothesis for A does involve B parameters.
- Type I hypotheses are functions of the cell counts for unbalanced data; the hypotheses are not usually the same hypotheses that are tested if the data are balanced.
- Type I sums of squares are useful for polynomial models where you want to know the contribution of a term as though it had been made orthogonal to preceding effects. Thus, in polynomial models, Type I sums of squares correspond to tests of the orthogonal polynomial effects.

Type II

The Type II tests can also be calculated by comparing the error sums of squares (SS) for subset models. The Type II SS are the reduction in error SS due to adding the term after all other terms have been added to the model except terms that contain the effect being tested. An effect is contained in another effect if it can be derived by deleting variables from the latter effect. For example, A and B are both contained in A*B. For this model

Source	Type I SS
A	$SS(A \mid \mu, B)$
B	$SS(B \mid \mu, A)$
A*B	$SS(A * B \mid \mu, A, B)$

Type II SS have these properties:

- Type II SS do not necessarily sum to the model SS.
- The hypothesis for an effect does not involve parameters of other effects except for containing effects (which it must involve to be estimable).
- Type II SS are invariant to the ordering of effects in the model.
- For unbalanced designs, Type II hypotheses for effects that are contained in other effects are not usually the same hypotheses that are tested if the data are balanced. The hypotheses are generally functions of the cell counts.

Type III and Type IV

Type III and Type IV sums of squares (SS), sometimes referred to as partial sums of squares, are considered by many to be the most desirable; see Searle (1987, Section 4.6). These SS cannot, in general, be computed by comparing model SS from several models using PROC GLM's parameterization. However, they can sometimes be computed by reduction for methods that reparameterize to full rank, when such a reparameterization effectively imposes Type III linear constraints on the parameters. In PROC GLM, they are computed by constructing a hypothesis matrix L and then computing the SS associated with the hypothesis $L\beta = 0$. As long as there are no missing cells in the design, Type III and Type IV SS are the same.

These are properties of Type III and Type IV SS:

- The hypothesis for an effect does not involve parameters of other effects except for containing effects (which it must involve to be estimable).
- The hypotheses to be tested are invariant to the ordering of effects in the model.
- The hypotheses are the same hypotheses that are tested if there are no missing cells. They are not functions of cell counts.
- The SS do not generally add up to the model SS and, in some cases, can exceed the model SS.

The SS are constructed from the general form of estimable functions. Type III and Type IV tests are different only if the design has missing cells. In this case, the Type III tests have an orthogonality property, while the Type IV tests have a balancing property.

7.3 Additional ANOVA Models

7.3.1 Randomized Block Designs (BRBD)

As in the one-way ANOVA one is primarily interested in potential differences in the treatment means as given by,

$$y_{ij} = \mu_i + e_{ij} = \mu + \tau_i + e_{ij},$$

for $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, n$, $N = na$. In some cases the differences in the treatment means (or $\tau_i's$) is hidden by excessive variation within the treatment groups, in which case, there may be another variable (called a block or blocking effect) that can "explain" this variation by partitioning the variance into smaller homogenous parts. Models of this type are called Balanced Randomized Block Design.The model becomes,

$$y_{ijk} = \mu + \tau_i + \beta_j + e_{ijk},$$

for $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$ and $k = 1, 2, \ldots, n$, $N = nab$.
Littell's text SAS for Linear Models contains examples for most of the models found in this chapter. You should look at his examples and reproduce his results using either SAS or R. In the interest of time I will compute the expected values of the mean square terms are for the various terms in the model so that one can form the correct ratios as test statistics.

Expected Mean Squares for the BRBD

The expected value for the mean square terms[4] are,

$$E(MS_{block}) = \sigma^2 + a\frac{\sum_{j=1}^{b}\beta_j^2}{(b-1)},$$

[4] Assuming that both the treatment and block effects are fixed.

$$E(MS_{treatment}) = \sigma^2 + b\frac{\sum_{i=1}^{a}\tau_i^2}{(a-1)},$$

and

$$E(MS_{error}) = \sigma^2.$$

7.3.2 Latin Squares Model

Suppose that one wants to incorporate another blocking factor without imposing the added cost associated with having the additional sample size. A method of doing this is called the Latin Squares model[5]. The model is described as;

$$y_{ijk} = \mu + \rho_i + \kappa_j + \tau_k + e_{ijk},$$

for $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, r$ and $k = 1, 2, \ldots, r$, $N = r^2$. The two blocking effects are specified by ρ_i and κ_j and the treatment effect is given by τ_k.

Expected Mean Squares for the Latin Square

The expected value for the mean square terms are,

$$E(MS_{row}) = \sigma^2 + r\frac{\sum_{i=1}^{r}\rho_i^2}{(r-1)},$$

$$E(MS_{column}) = \sigma^2 + r\frac{\sum_{j=1}^{r}\kappa_j^2}{(r-1)},$$

$$E(MS_{block}) = \sigma^2 + r\frac{\sum_{k=1}^{r}\tau_k^2}{(r-1)},$$

and

$$E(MS_{error}) = \sigma^2.$$

7.3.3 Two-Way Factorial Design

Suppose that one has two treatment effects, A (with a levels) and B (with b levels), then there are $a \times b$ possible treatment combinations. If this model is repeated r times then the model is said to be a Balanced Two-way Factorial design model. If some of the $a \times b$ cells contain an unequal number of observations then the model is said to be unbalanced.[6]

The two-way factorial model is given by,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk},$$

for $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$ and $k = 1, 2, \ldots, r$, $N = abr$. The two treatment effects are specified by α_i and β_j and the interaction (multiplicative) term is given by γ_{ij}. This term is often written as $\gamma_{ij} = (\alpha\beta)_{ij}$.

[5]The design of the Latin Squares is an interesting topic (for some people - of which I am not one) which you can find in separate texts on the Design of Experiments.

[6]The unbalanced design is complicated and will be treated as a separate topic.

Expected Mean Squares for the Balanced Two-way with Interaction

The expected value for the mean square terms are,

$$E(MS_A) = \sigma^2 + rb\frac{\sum_{i=1}^{a}\alpha_i^2}{(a-1)},$$

$$E(MS_B) = \sigma^2 + ra\frac{\sum_{j=1}^{b}\beta_j^2}{(b-1)},$$

$$E(MS_{AB}) = \sigma^2 + r\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\gamma_{ij}^2}{(a-1)(b-1)},$$

and

$$E(MS_E) = \sigma^2.$$

7.3.4 Nested Balanced Models

This model differs from the two-way factorial model in that the second factor, B, is nested within treatment A. This nesting creates a dependency upon a particular level of A and does not allow one to estimate the interaction effects in the model. The model is

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + e_{ijk}$$

for $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$ and $k = 1, 2, \ldots, n$, $N = abn$. The two treatment effects are specified by α_i and β_j where β_j is dependent upon the treatment i.

Expected Mean Squares for the Balance Nested with Fixed Effects

The expected value for the mean square terms are,

$$E(MS_A) = \sigma^2 + nb\frac{\sum_{i=1}^{a}\alpha_i^2}{(a-1)},$$

$$E(MS_{B(A)}) = \sigma^2 + r\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\beta_{j(i)}^2}{a(b-1)},$$

and

$$E(MS_E) = \sigma^2.$$

7.4 Balanced Random Effects Models

In this section, the treatment or block effects are no longer considered to be fixed effects. Instead, they denote effects that are randomly selected from a larger population of effects. That is, in the fixed effects model the a treatment groups were denoted as $\tau_1, \tau_2, \ldots, \tau_a$ with a general hypothesis of interest being $H_0 : \tau_i = \tau_{i'}$ for all pairs, i and i'. Now these treatments are considered to be randomly selected from a normal population with mean zero and unknown variance, σ_A^2. That is, $\tau_i \sim N(0, \sigma_A^2)$. The inference of interest concerns the unknown variances given by, $H_0 : \sigma_A^2 = 0$.

7.4.1 One Way Model with Random Treatments

Expected Mean Squares for the Balance One Way Model

The expected value for the mean square terms are,

$$E(MS_A) = \sigma^2 + n\sigma_A^2,$$

and

$$E(MS_E) = \sigma^2.$$

The estimation of the variances using the least squares estimates[7] follows by setting the Mean Square for the treatment equal to it expected value and solving. That is,

$$\hat{\sigma}^2 = MS_E$$

and

$$\hat{\sigma}_A^2 = \frac{(MS_A - MS_E)}{n}.$$

7.4.2 Randomized Block Designs (BRBD)

The Balanced Randomized Block Design model becomes,

$$y_{ijk} = \mu + \tau_i + \beta_j + e_{ijk},$$

for $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$ and $k = 1, 2, \ldots, n$, $N = nab$. In this model one can assume that the blocks are randomly selected, that is, $\beta_j \sim N(0, \sigma_B^2)$. In this case, the analysis is exactly the same as the fixed effects model since one is not usually interested in the blocking effect and it has been added to the model to account for unexplained variability in the response variable. Hence, one expects that $\sigma_B^2 > 0$.
Suppose the block and treatment effects are random, that is, $\tau_i \sim N(0, \sigma_A^2)$ and that the blocks are independent of the treatment then the following holds.

Expected Mean Squares for the BRBD with Random Blocks and Treatments

The expected value for the mean square terms are,

$$E(MS_B) = \sigma^2 + na\sigma_B^2,$$

$$E(MS_A) = \sigma^2 + nb\sigma_A^2,$$

and

$$E(MS_E) = \sigma^2.$$

The estimation of the variances using the least squares estimates[8] follows by setting the Mean Square for the treatment equal to it expected value and solving. That is,

$$\hat{\sigma}^2 = MS_E,$$

$$\hat{\sigma}_B^2 = \frac{(MS_B - MS_E)}{na},$$

and

$$\hat{\sigma}_A^2 = \frac{(MS_A - MS_E)}{nb}.$$

[7] (PROC GLM), Since all the terms in the random model are variances there are better ways to estimate these variances than using the least squares approach. PROC MIXED uses a more powerful likelihood approach when estimating variances.
[8] (PROC GLM)

7.4.3 Two-Way Factorial Design

The two-way factorial model is given by,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk},$$

for $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$ and $k = 1, 2, \ldots, r$, $N = abr$. Assume that the two treatment effects are specified by α_i and β_j and the interaction term is given by γ_{ij} are all random.

Expected Mean Squares for the Balanced Two-way with Interaction – Both Effects Random

The expected value for the mean square terms are,

$$E(MS_A) = \sigma^2 + rb\sigma_A^2 + r\sigma_{AB}^2,$$
$$E(MS_B) = \sigma^2 + ra\sigma_B^2 + r\sigma_{AB}^2,$$
$$E(MS_{AB}) = \sigma^2 + r\sigma_{AB}^2,$$

and

$$E(MS_E) = \sigma^2.$$

Estimation of Variances

$$\hat{\sigma}_E^2 = MS_E,$$
$$\hat{\sigma}_{AB}^2 = \frac{(MS_{AB} - MS_E)}{r},$$
$$\hat{\sigma}_B^2 = \frac{(MS_B - MS_{AB})}{ra},$$

and

$$\hat{\sigma}_A^2 = \frac{(MS_A - MS_{AB})}{rb}.$$

7.4.4 Nested Balanced Models

The model is

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + e_{ijk}$$

for $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$ and $k = 1, 2, \ldots, n$, $N = abn$. The two treatment effects are specified by α_i and β_j where β_j is dependent upon the treatment i. As seen in this next section, the expected mean square is dependent upon the nature of the model, that is, whether or not the effects are fixed, random, or mixed.

Expected Mean Squares for the Balance Nested

Both Effects are Fixed

The expected value for the mean square terms are,

$$E(MS_A) = \sigma^2 + nb\frac{\sum_{i=1}^{a} \alpha_i^2}{(a-1)},$$

$$E(MS_{B(A)}) = \sigma^2 + r\frac{\sum_{i=1}^{a}\sum_{j=1}^{b} \beta_{j(i)}^2}{a(b-1)},$$

and

$$E(MS_E) = \sigma^2.$$

A Fixed, B random

The expected value for the mean square terms are,

$$E(MS_A) = \sigma^2 + n\sigma^2_{B(A)} + nb\frac{\sum_{i=1}^{a}\alpha_i^2}{(a-1)},$$

$$E(MS_{B(A)}) = \sigma^2 + n\sigma^2_{B(A)},$$

and

$$E(MS_E) = \sigma^2.$$

The test statistics for the fixed main effects is $\frac{MS_A}{MS_{B(A)}}$ and the estimation of the variance for $B(A)$ is,

$$\hat{\sigma}^2_{B(A)} = \frac{(MS_{B(A)} - MS_E)}{n}.$$

Both Effects are Random

The expected value for the mean square terms are,

$$E(MS_A) = \sigma^2 + n\sigma^2_{B(A)} + nb\sigma^2_A,$$

$$E(MS_{B(A)}) = \sigma^2 + n\sigma^2_{B(A)},$$

and

$$E(MS_E) = \sigma^2.$$

The estimation of the variances for A and $B(A)$ is,

$$\hat{\sigma}^2_A = \frac{(MS_A - MS_{B(A)})}{nb},$$

and

$$\hat{\sigma}^2_{B(A)} = \frac{(MS_{B(A)} - MS_E)}{n}.$$

Chapter 8

Multiple Comparisons

8.1 Basic Concepts

8.1.1 Error Rates

Definition: 8.1 (Comparisonwise Error Rate (CER)) *The comparison error rate is given by*

$$CER = \Pr[\text{Interval does not contain the parameter}]$$

when computing confidence intervals. It is given by

$$\Pr[\text{Reject } H_0 \mid H_0 \text{ is true}]$$

when testing a null hypothesis, H_0. Note this error is the actual type I error when making a single comparison.

Definition: 8.2 (Familywise Error Rate (FWE)) *The familywise error rate is the probability of making a false claim when the entire family of inference is considered. It is given by*

$$FWE = \Pr[\text{at least one interval is incorrect}] = 1 - \Pr[\text{all intervals are correct}]$$

when considering confidence intervals and is

$$FWE = \Pr[\text{reject at least one hypothesis} \mid \text{all hypotheses are true}]$$

when considering tests of hypothesis.

Example

Suppose that one has $k = 20$ tests of independent hypotheses for which the $CER = .05$ and that $m = 8$ of these are true null hypotheses, then the $FWE = 1 - (0.95)^8 = 33.7\%$. If all 20 hypotheses are true then $FWE = 64.2\%$.

Definition: 8.3 (False Discovery Rate (FDR)) *The False Discovery Rate is the expected proportion of erroneously rejected null hypotheses among the rejected null hypotheses. Let R = number of hypotheses rejected and V = number of erroneously rejected hypotheses (unknown), then*

$$FDR = E[V/R].$$

8.1.2 Single-Step Tests

Suppose that one has k hypotheses with associated individual p-vales $p_1, \ldots, p_k$, with the corresponding order,

$$p_{(1)} \leq p_{(2)}, \ldots, \leq p_{(k)}.$$

Bonferroni and Sidak Methods

Note if one has k hypotheses and wished to control the overall FWE at α then the individual CER level need to by $CER = 1 - (1-\alpha)^{1/k}$. This method is known as the Sidak Method. Another method is given by Bonferroni which makes use of the Bonferroni inequality where the $CER = FWE/k$. That is, Bonferroni rejects

$$H_i \quad \text{if} \quad p_i \leq \alpha/k$$

and Sidak rejects

$$H_i \quad \text{if} \quad p_i \leq 1 - (1-\alpha)^{1/k}.$$

The adjusted p-value for any hypothesis is the smallest FWE at which the hypothesis would be rejected. The Bonferroni adjusted p-values are

$$\begin{aligned} \tilde{p}_i &= kp_i \text{ when } kp_i \leq 1 \\ &= 0 \text{ otherwise} \end{aligned}$$

The Sidak adjusted p-values are

$$\tilde{p}_i = 1 - (1-p_i)^k.$$

Simes' Method

Simes modified the Bonferroni method where the test for the global H_0 is based upon the individual test p-values. That is, reject H_0 if $p_{(j)} \leq j\alpha/k$ for at least one value j. The adjusted p-value is

$$\widetilde{p_{SIM}} = k \; min(p_{(1)}, p_{(2)}/2. \ldots, p_{(k)}/k).$$

Since the adjusted Bonferroni p-value $= k \; p_{(1)} \geq \widetilde{p_{SIM}}$ Simes global test is uniformly more powerful than the Bonferroni global test.

8.1.3 Sequentially Rejective Methods

The previous methods were called single step methods as only a single step is needed in order to determine the appropriate critical value for all the tests or confidence intervals. Sequentially rejective methods are stepwise procedures which differ in that the result of a given test depends on the results of the other tests.

Bonferroni-Holm Method

Assume that there are k hypotheses of interest with corresponding p-values, p_i, $i = 1, 2, \ldots, k$. The single step Bonferroni method rejects the hypothesis, H_i if $p_i \leq \alpha/k$, where α is the FWE level. Holm modified this method by ordering the p-values such that

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k-1)} \leq p_{(k)}$$

and one rejects the hypothesis corresponding to the smallest p-value, give by, $H_{(1)}$ if $p_{(1)} \leq \alpha/k$ otherwise all the hypotheses are retained (not rejected). Then one considers the next hypothesis (corresponding to $H_{(2)}$) where this is rejected if $p_{(2)} \leq \alpha/(k-1)$ otherwise the remaining hypotheses are retained. This procedure

continues to where the hypothesis corresponding to the j^{th} smallest p-value is rejected if, $p_{(i)} \leq \alpha/(k-i)$. The adjusted p-value is given by

$$\tilde{p}_{(i)} = max(\tilde{p}_{(i-1)}, (k-i+1)p_{(i)}).$$

Sidak-Holm Method

A similar method is given using the Sidak equation rather than the Bonferroni. That is,

$$\tilde{p}_{(i)} = max(\tilde{p}_{(i-1)}, 1-(1-p_{(i)})^{(k-i+1)}).$$

Hochberg's Method – A Step-Up Test

The preceding tests are step-down procedures since you begin with the smallest p-value (most significant test) and step down to the higher p-values (less significant tests). The step-up test works in the opposite direction and is based upon Simes' method. One starts with the largest p-value (least significant test), given by $p_{(k)}$ where all hypotheses are rejected if $p_{(k)} \leq \alpha$, otherwise hypothesis H_k is retained, then one considers the next largest p-value where all remaining hypotheses are rejected if $p_{(k)} \leq \alpha/2$, otherwise hypothesis $H_{(k-1)}$ is retained. The next hypothesis is tested using $\alpha/3$ and so forth. The adjusted p-values are given by

$$\tilde{p}_{(k-i)} = min(\tilde{p}_{(k-i+1)}, (i+1)p_{(k-i+1)}).$$

8.2 Multiple Comparisons in the Linear Model

In this chapter we consider the problem of making multiple comparisons in the general linear model. This model can be written as

$$\vec{y} = X\vec{\beta} + \vec{e}.$$

Recall, that since X is not full column rank (in the ANOVA setting) then one uses the non-unique solution to the normal equations given by,

$$\tilde{\beta} = (X'X)^{-}X'y$$

where $(X'X)^{-}$ is the generalized inverse of $(X'X)$ satisfying

$$(X'X)(X'X)^{-}(X'X) = (X'X).$$

The least squares estimate for σ^2 is given by

$$\hat{\sigma}^2 = SS_E/(N - rank(X))$$

where

$$SS_E = (y - X\tilde{\beta})'(y - X\tilde{\beta}) = y'(I-H)y = y'y - \tilde{\beta}'X'y.$$

Since, $\tilde{\beta}$ is not unique, one restricted oneself to estimable functions of β, say $c'\beta$ from which one has

$$Var(c'\tilde{\beta}) = \sigma^2 c'(X'X)^{-}c$$

and

$$s.e.(c'\tilde{\beta}) = \hat{\sigma}\sqrt{c'(X'X)^{-}c}.$$

The simultaneous confidence intervals can now be written as

$$c'\tilde{\beta} \pm c_\alpha \; s.e.(c'\tilde{\beta}),$$

where c_α is the appropriate critical point. For individual intervals with CER = α, $c_\alpha = t_{\alpha/2}(df = N - rank(X))$, the conservative Bonferroni intervals would have $c_\alpha = t_{\alpha/2k}(df = N - rank(X))$, k = number of intervals for FEW = α.

8.3 Adjustments to the p-Value

Suppose you test $H_{01}, \ldots, H_{0m}$, and obtain the p-values $p_1, \ldots, p_m$. Denote the ordered p-values as $p_{(1)} \leq \ldots \leq p_{(m)}$ and order the tests appropriately: $H_{0(1)}, \ldots, H_{0(m)}$. Suppose you know m_0 of the null hypotheses are true and $m_1 = m - m_0$ are false. Let R indicate the number of null hypotheses rejected by the tests, where V of these are incorrectly rejected (that is, V tests are Type I errors) and $R - V$ are correctly rejected (so $m_1 - R + V$ tests are Type II errors). This information is summarized in the following table:

	Null is Rejected	Null is Not Rejected	Total
Null is True	V	$m_0 - V$	m_0
Null is False	$R - V$	$m_1 - R + V$	m_1
Total	R	$m - R$	m

The **familywise error rate (FWE)** is the overall Type I error rate for all the comparisons (possibly under some restrictions); that is, it is the maximum probability of incorrectly rejecting one or more null hypotheses:

$$FWE = \Pr[V > 0].$$

The FWE is also known as the *maximum experiment-wise error rate (MEER)*.

The **false discovery rate (FDR)** is the expected proportion of incorrectly rejected hypotheses among all rejected hypotheses:

$$FDR = E(\frac{V}{R} \mid R > 0) \times \Pr[R > 0].$$

Under the overall null hypothesis (all the null hypotheses are true), the $FDR = FWE$ since $V = R$ $E\left(\frac{V}{R}\right) = 1 \times \Pr\left(\frac{V}{R} = 1\right) = \Pr(V > 0)$. Otherwise, FDR is always less than FWE, and an FDR-controlling adjustment also controls the FWE. Another definition used is the *positive false discovery rate*:

$$pFDR = E(\frac{V}{R} \mid R > 0).$$

The p-value adjustment methods discussed in the following sections attempt to correct the raw p-values while controlling either the FWE or the FDR. Note that the methods might impose some restrictions in order to achieve this; restrictions are discussed along with the methods in the following sections. Discussions and comparisons of some of these methods are given in Dmitrienko et. al. (2005), Dudoit, Shaffer, and Boldrick (2003), Westfall et. al. (1999), and Brown and Russell (1997).

Familywise Error Rate Controlling Adjustments

PROC MULTTEST provides several p-value adjustments to control the familywise error rate. Single-step adjustment methods are computed without reference to the other hypothesis tests under consideration. The available single-step methods are the Bonferroni and Sidak adjustments, which are simple functions of the raw p-values that try to distribute the significance level α across all the tests, and the bootstrap and permutation resampling adjustments, which require the raw data. The Bonferroni and Sidak methods are calculated from the permutation distributions when exact permutation tests are used with the CA or Peto test.

Stepwise tests, or sequentially rejective tests, order the hypotheses in step-up (least significant to most significant) or step-down fashion, then sequentially determine acceptance or rejection of the nulls. These tests are more powerful than the single-step tests, and they do not always require you to perform every test. However, PROC MULTTEST still adjusts every p-value. PROC MULTTEST provides the following stepwise p-value adjustments: step-down Bonferroni (Holm), step-down Sidak step-down bootstrap and permutation

resampling, Hochberg (1988) step-up, Hommel (1988) Fisher's combination method, and the Stouffer-Liptak combination method. Adaptive versions of Holm's step-down Bonferroni and Hochberg's step-up Bonferroni methods, which require an estimate of the number of true null hypotheses, are also available.

Liu (1996) shows that all single-step and stepwise tests based on marginal p-values can be used to construct a closed test Marcus, Peritz, and Gabriel, (1976) Dmitrienko et .al. (2005)). Closed testing methods not only control the familywise error rate at size α, but are also more powerful than the tests on which they are based. Westfall and Wolfinger (2000) note that several of the methods available in PROC MULTTEST are closed – namely, the step-down methods, Hommel's method, and Fisher's combination; see that reference for conditions and exceptions.

All methods except the resampling methods are calculated by simple functions of the raw p-values or marginal permutation distributions; the permutation and bootstrap adjustments require the raw data. Because the resampling techniques incorporate distributional and correlational structures, they tend to be less conservative than the other methods.

When a resampling (bootstrap or permutation) method is used with only one test, the adjusted p-value is the bootstrap or permutation p-value for that test, with no adjustment for multiplicity, as described by Westfall and Soper (1994).

Bonferroni

The Bonferroni adjusted p-value for test $i, i = 1, \ldots, m$ is simply $\tilde{p}_i = mp_i$. If the adjusted p-value exceeds 1, it is set to 1. The Bonferroni test is conservative but always controls the familywise error rate.

If the unadjusted p-values are computed by using exact permutation distributions, then the Bonferroni adjustment for p_i is $\tilde{p}_i = p_1^* + \cdots + p_m^*$, where p_j^* is the largest p-value from the permutation distribution of test satisfying $p_j^* \leq p_i$, or 0 if all permutational p-values of test are greater than p_i. These adjustments are much less conservative than the ordinary Bonferroni adjustments because they incorporate the discrete distributional characteristics. However, they remain conservative in that they do not incorporate correlation structures between multiple contrasts and multiple variables Westfall and Wolfinger (1997).

Sidak

A technique slightly less conservative than Bonferroni is the SIdak p-value (SIdak (1967)), which is $\tilde{p}_i = 1 - (1-p_i)^m$. It is exact when all of the p-values are uniformly distributed and independent, and it is conservative when the test statistics satisfy the positive orthant dependence condition Holland and Copenhaver (1987).

If the unadjusted p-values are computed by using exact permutation distributions, then the Sidak adjustment for p_i is $\tilde{p}_i = 1 - (1 - p_1^*) \cdots (1 - p_m^*)$, where the p_j^* are as described previously. These adjustments are less conservative than the corresponding Bonferroni adjustments, but they do not incorporate correlation structures between multiple contrasts and multiple variables Westfall and Wolfinger (1997).

Bootstrap

The bootstrap method creates pseudo-data sets by sampling observations with replacement from each within-stratum pool of observations. An entire data set is thus created, and p-values for all tests are computed on this pseudo-data set. A counter records whether the minimum p-value from the pseudo-data set is less than or equal to the actual p-value for each base test. (If there are m tests, then there are m such counters.) This process is repeated a large number of times, and the proportion of resampled data sets where the minimum pseudo p-value is less than or equal to an actual p-value is the adjusted p-value reported by PROC MULTTEST. The algorithms are described in Westfall and Young (1993).

In the case of continuous data, the pooling of the groups is not likely to re-create the shape of the null hypothesis distribution, since the pooled data are likely to be multimodal. For this reason, PROC MULTTEST

automatically mean-centers all continuous variables prior to resampling. Such mean-centering is akin to resampling residuals in a regression analysis, as discussed by Freedman (1981). You can specify the NOCENTER option if you do not want to center the data.

The bootstrap method implicitly incorporates all sources of correlation, from both the multiple contrasts and the multivariate structure. The adjusted p-values incorporate all correlations and distributional characteristics. This method always provides weak control of the familywise error rate, and it provides strong control when the subset pivotality condition holds; that is, for any subset of the null hypotheses, the joint distribution of the p-values for the subset is identical to that under the complete null Westfall and Young (1993).

Permutation

The permutation-style-adjusted p-values are computed in identical fashion as the bootstrap adjusted p-values, with the exception that the within-stratum resampling is performed without replacement instead of with replacement. This produces a rerandomization analysis such as in Brown and Fears (1981) and Heyse and Rom (1988). In the spirit of rerandomization analyses, the continuous variables are not centered prior to resampling. This default can be overridden by using the CENTER option.

The permutation method implicitly incorporates all sources of correlation, from both the multiple contrasts and the multivariate structure. The adjusted p-values incorporate all correlations and distributional characteristics. This method always provides weak control of the familywise error rate, and it provides strong control of the familywise error rate under the subset pivotality condition, as described in the preceding section.

Step-Down Methods

Step-down testing is available for the Bonferroni, Sidak, bootstrap, and permutation methods. The benefit of using step-down methods is that the tests are made more powerful (smaller adjusted p-values) while, in most cases, maintaining strong control of the familywise error rate. The step-down method was pioneered by Holm (1979) and further developed by Shaffer (1986), Holland and Copenhaver (1987), and Hochberg and Tamhane (1987).

The Bonferroni step-down (Holm) p-values $\tilde{p}_{(1)}, \ldots, \tilde{p}_{(m)}$ are obtained from

$$\tilde{p}_{(i)} = \begin{cases} mp_{(1)} & \text{for } i = 1 \\ \max(\tilde{p}_{(i-1)}, (m-i+1)p_{(i)}) & \text{for } i = 2, \ldots, m \end{cases}$$

As always, if any adjusted p-value exceeds 1, it is set to 1. The Sidak step-down p-values are determined similarly:

$$\tilde{p}_{(i)} = \begin{cases} 1-(1-p_{(1)})^m & \text{for } i = 1 \\ \max\left(\tilde{p}_{(i-1)}, 1-(1-p_{(i)})^{m-i+1}\right) & \text{for } i = 2, \ldots, m \end{cases}$$

Step-down Bonferroni adjustments that use exact tests are defined as

$$\tilde{p}_{(i)} = \begin{cases} p^*_{(1)} + \cdots + p^*_{(m)} & \text{for } i = 1 \\ \max\left(\tilde{p}_{(i-1)}, p^*_{(i)} + \cdots + p^*_{(m)},\right) & \text{for } i = 2, \ldots, m \end{cases}$$

where the p^*_j are defined as before. Note that p^*_j is taken from the permutation distribution corresponding to the j^{th}-smallest unadjusted p-value. Also, any $\tilde{p}_j$ greater than 1.0 is reduced to 1.0.

Step-down Sidak adjustments for exact tests are defined analogously by substituting $1-(1-p^*_{(j)})\cdots(1-p^*_{(m)})$ for $p^*_{(j)} + \cdots + p^*_{(m)}$.

The resampling-style step-down methods are analogous to the preceding step-down methods; the most extreme p-value is adjusted according to all m tests, the second-most extreme p-value is adjusted according to $(m-1)$ tests, and so on. The difference is that all correlational and distributional characteristics are incorporated when you use resampling methods. More specifically, assuming the same ordering of p-values as discussed previously, the resampling-style step-down-adjusted p-value for test I is the probability that the minimum pseudo-p-value of tests $i, \ldots, m$ is less than or equal to p_i.

This probability is evaluated by using Monte Carlo simulation, as are the previously described resampling-style-adjusted p-values. In fact, the computations for step-down-adjusted p-values are essentially no more time-consuming than the computations for the non-step-down-adjusted p-values. After Monte Carlo, the step-down-adjusted p-values are corrected to ensure monotonicity; this correction leaves the first adjusted p-values alone, then corrects the remaining ones as needed. The step-down method approximately controls the family wise error rate, and it is described in more detail by Westfall and Young (1993), Westfall et. al. (1999), and Westfall and Wolfinger (2000).

Hommel

Homme‘s (1988) method is a closed testing procedure based on Simes‘s; test (Simes, 1986). The Simes p-value for a joint test of any set of S hypotheses with p-values $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(S)}$ is

$$\min((S/1)p_{(1)}, (S/2)p_{(2)}, \ldots, (S/S)p_{(S)}).$$

The Hommel-adjusted p-value for test is the maximum of all such Simes' p-values, taken over all joint tests that include j as one of their components.

Hochberg-adjusted p-values are always as large or larger than Hommel-adjusted p-values. Sarkar and Chang (1997) shows that Simes‘ method is valid under independent or positively dependent p-values, so Hommel‘s and Hochberg‘s methods are also valid in such cases by the closure principle.

Hochberg

Assuming p-values are independent and uniformly distributed under their respective null hypotheses, Hochberg (1988) demonstrates that Holm‘s step-down adjustments control the familywise error rate even when calculated in step-up fashion. Since the adjusted p-values are uniformly smaller for Hochberg‘s method than for Holm‘s method, the Hochberg method is more powerful. However, this improved power comes at the cost of having to make the assumption of independence. Hochberg‘s method can be derived from Hommel‘s (Liu, 1996), and is thus also derived from Simes‘ test (Simes, 1986).

Hochberg-adjusted p-values are always as large or larger than Hommel-adjusted p-values. Sarkar and Chang (1997) showed that Simes‘ method is valid under independent or positively dependent p-values, so Hommel‘s and Hochberg‘s methods are also valid in such cases by the closure principle.

The Hochberg-adjusted p-values are defined in reverse order of the step-down Bonferroni:

$$\tilde{p}_{(i)} = \begin{cases} p_{(m)} & \text{for } i = m \\ \min\left(\tilde{p}_{(i+1)}, (m-i+1)p_{(i)}\right) & \text{for } i = m-1, \ldots, 1 \end{cases}$$

Fisher Combination

FISHER C option requests adjusted p-values by using closed tests, based on the idea of Fisher‘s combination test. The Fisher combination test for a joint test of any set of S hypotheses with p-values uses the chi-square statistic $\chi^2 = -2\sum \log(p_i)$, with $2S$ degrees of freedom. The FISHER C adjusted p-value for test j is the maximum of all p-values for the combination tests, taken over all joint tests that include j as one of their components. Independence of p-values is required for the validity of this method.

Stouffer-Liptak Combination

STOUFFER option requests adjusted p-values by using closed tests, based on the Stouffer-Liptak combination test. The Stouffer combination joint test of any set of S one-sided hypotheses with p-values, $p_1, \ldots, p_S$, yields the p-value, $1 - \Phi\left(\frac{1}{\sqrt{S}} \sum_i \Phi^{-1}(1-p_i)\right)$. The STOUFFER adjusted p-value for test j is the maximum of all p-values for the combination tests, taken over all joint tests that include j as one of their components.

Independence of the one-sided p-values is required for the validity of this method. Westfall (2005) shows that the Stouffer-Liptak adjustment might have more power than the Fisher combination and Simes' adjustments when the test results reinforce each other.

Adaptive Adjustments

Adaptive adjustments modify the FWE- and FDR-controlling procedures by taking an estimate of the number m_0 or proportion π_0 of true null hypotheses into account. The adjusted p-values for Holm's and Hochberg's methods involve the number of unadjusted p-values larger than (i), $m - i + 1$. So the minimal significance level at which the i^{th} ordered p-value is rejected implies that the number of true null hypotheses is $m - i + 1$. However, if you know m_0, then you can replace $m - i + 1$ with $\min(m_0, m - i + 1)$, thereby obtaining more power while maintaining the original α-level significance.

Since m_0 is unknown, there are several methods used to estimate the value – see the NTRUENULL= option for more information. The estimation method described by Hochberg and Benjamini (1990) considers the graph of $1 - p_{(i)}$ versus i, where the $p_{(i)}$ are the ordered p-values of your tests. See Output 61.6.4 for an example. If all null hypotheses are actually true ($m_0 = m$), then the p-values behave like a sample from a uniform distribution and this graph should be a straight line through the origin. However, if points in the upper-right corner of this plot do not follow the initial trend, then some of these null hypotheses are probably false and $0 < m_0 < m$.

The ADAPTIVEHOLM option uses this estimate of m_0 to adjust the step-up Bonferroni method while the ADAPTIVEHOCHBERG option adjusts the step-down Bonferroni method. Both of these methods are due to Hochberg and Benjamini (1990). When m_0 is known, these procedures control the familywise error rate in the same manner as their nonadaptive versions but with more power; however, since m_0 must be estimated, the FWE control is only approximate. The ADAPTIVEFDR and PFDR options also use $\hat{m}_0$, and are described in the following section.

The adjusted p-values for the ADAPTIVEHOLM method are computed by

$$\tilde{p}_{(i)} = \begin{cases} \min(m, \hat{m}_0) p_{(1)} & \text{for } i = 1 \\ \max\left[\tilde{p}_{(i-1)}, \min(m - i + 1, \hat{m}_0) p_{(i)}\right] & \text{for } i = 2, \ldots, m \end{cases}$$

The adjusted p-values for the ADAPTIVEHOCHBERG method are computed by

$$\tilde{p}_{(i)} = \begin{cases} \min(1, \hat{m}_0) p_{(m)} & \text{for } i = m \\ \min\left[\tilde{p}_{(i+1)}, \min(m - i + 1, \hat{m}_0) p_{(i)}\right] & \text{for } i = m - 1, \ldots, 1 \end{cases}$$

False Discovery Rate Controlling Adjustments

Methods that control the false discovery rate(FDR) were described by Benjamini and Hochberg (1995). These adjustments do not necessarily control the familywise error rate (FWE). However, FDR-controlling methods are more powerful and more liberal, and hence reject more null hypotheses, than adjustments protecting the FWE. FDR-controlling methods are often used when you have a large number of null hypotheses. To control the FDR, Benjamini and Hochberg's (1995) linear step-up method is provided, as well as an adaptive

version, a dependence version, and bootstrap and permutation resampling versions. Storey's (2000) pFDR methods are also provided.

The FDR option requests p-values that control the false discovery rate described by Benjamini and Hochberg (1995). These linear step-up adjustments are potentially much less conservative than the Hochberg adjustments.

The FDR-adjusted p-values are defined in step-up fashion, like the Hochberg adjustments, but with less conservative multipliers:

$$\tilde{p}_{(i)} = \begin{cases} p_{(m)} & \text{for } i = m \\ \min\left(\tilde{p}_{(i+1)}, \frac{m}{i} p_{(i)}\right) & \text{for } i = m-1, \ldots, 1 \end{cases}$$

The FDR method is guaranteed to control the false discovery rate at level $\leq \frac{m_0}{m}\alpha \leq \alpha$ when you have independent p-values that are uniformly distributed under their respective null hypotheses. Benjamini and Yekateuli (2001) show that the false discovery rate is also controlled at level $\leq \frac{m_0}{m}\alpha$ when the *positive regression dependent* condition holds on the set of the true null hypotheses, and they provide several examples where this condition is true.

The positive regression dependent condition on the set of the true null hypotheses holds if the joint distribution of the test statistics $X = (X_1, \ldots, X_m)$ for the null hypotheses $H_{01}, \ldots, H_{0m}$ satisfies: $\Pr(X \in A | X_i = x)$ is nondecreasing in x for each X_i where H_{0i} is true, for any increasing set A. The set A is increasing if $x \in A$ and $y \geq x$ implies $y \in A$

Dependent False Discovery Rate

The DEPENDENTFDR option requests a false discovery rate controlling method that is always valid for p-values under any kind of dependency (Benjamini and Yekateuli, 2001), but is thus quite conservative. Let $\gamma = \sum_{i=1}^{m} \frac{1}{i}$. The DEPENDENTFDR procedure always controls the false discovery rate at level $\leq \frac{m_0}{m}\alpha\gamma$. The adjusted p-values are computed as

$$\tilde{p}_{(i)} = \begin{cases} \gamma p_{(m)} & \text{for } i = m \\ \min\left(\tilde{p}_{(i+1)}, \gamma \frac{m}{i} p_{(i)}\right) & \text{for } i = m-1, \ldots, 1 \end{cases}$$

False Discovery Rate Resampling Adjustments

Bootstrap and permutation resampling methods to control the false discovery rate are available with the FDRBOOT and FDRPERM options (Yekateuli and Benjamini, 1999). These methods approximately control the false discovery rate when the *subset pivotality* condition holds, as discussed in the section Bootstrap, and when the p-values corresponding to the true null hypotheses are independent of those for the false null hypotheses.

The resampling methodology for the BOOTSTRAP and PERMUTATION methods is used to create B resamples. For the b^{th} resample, let $R^b(p)$ denote the number of p-values that are less than or equal to the observed p-value p. Let $r_\beta(p)$ be the $100(1-\beta)$ quantile of $\{R^1(p) \ldots R^b(p) \ldots R^B(p)\}$, and let $r(p)$ be the number of observed p-values less than or equal to p. Compute one of the following estimators:

- local estimator

$$Q_1(p) = \begin{cases} \frac{1}{B} \sum_{b=1}^{B} \frac{R^b(p)}{R^b(p) + r(p) - pm} & \text{if } r(p) - r_\beta(p) \geq pm \\ \#\{R^b(p) \geq 1\}/B & \text{otherwise} \end{cases}$$

- upper limit estimator

$$Q_\beta(p) = \begin{cases} \sup_{x\in[0,p]} \left(\frac{1}{B} \sum_{b=1}^{B} \frac{R^b(x)}{R^b(x) + r(x) - r_\beta(x)} \right) & \text{if } r(x) - r_\beta(x) \geq 0 \\ \#\{R^b(p) \geq 1\}/B & \text{otherwise} \end{cases}$$

where m is the number of tests and B is the number of resamples. Then for $Q = Q_1$ or Q_β, the adjusted p-values are computed as

$$\tilde{p}_{(i)} = \begin{cases} Q(p_{(m)}) & \text{for } i = m \\ \min\left(\tilde{p}_{(i+1)}, Q(p_{(i)})\right) & \text{for } i = m-1, \ldots, 1 \end{cases}$$

Adaptive False Discovery Rate

Since the FDR method controls the false discovery rate at $\leq \frac{m_0}{m}\alpha \leq \alpha$, knowledge of m_0 allows improvement of the power of the adjustment while still maintaining control of the false discovery rate. The ADAPTIVEFDR option requests adaptive adjusted p-values for approximate control of the false discovery rate, as discussed in Benjamini and Hochberg (2000). See the section Adaptive Adjustments for more details. These adaptive adjustments are also defined in step-up fashion but use an estimate $\hat{m}_0$ of the number of true null hypotheses:

$$\tilde{p}_{(i)} = \begin{cases} \frac{\hat{m}_0}{m} p_{(m)} & \text{for } i = m \\ \min\left(\tilde{p}_{(i+1)}, \frac{\hat{m}_0}{i} p_{(i)}\right) & \text{for } i = m-1, \ldots, 1 \end{cases}$$

Since $\hat{m}_0 \leq m$, the larger p-values are adjusted down. This means that, as defined, controlling the false discovery rate enables you to reject these tests at a level less than the observed p-value. However, by default, this reduction is prevented with an additional restriction: $\tilde{p}_{(i)} = \max\{\tilde{p}_{(i)}, p_{(i)}\}$.

To use this adjustment, Benjamini and Hochberg (2000) suggest first specifying the FDR option if at least one test is rejected at your level, then apply the ADAPTIVEFDR adjustment. Alternatively, Benjamini, Krieger, and Yekutieli (2006) apply the FDR adjustment at level $\frac{\alpha}{\alpha+1}$, then specify the resulting number of true hypotheses with the NTRUENULL= option and apply the ADAPTIVEFDR adjustment; they show that this —sl two-stage linear step-up procedure controls the false discovery rate at level α for independent test statistics.

Positive False Discovery Rate

The PFDR option computes the '$q-values$' $\hat{q}_\lambda(p_i)$ (Storey, 2002, Storey, Taylor, and Siegmund, 2004), which are adaptive adjusted p-values for strong control of the false discovery rate when the p-values corresponding to the true null hypotheses are independent and uniformly distributed. There are four versions of the PFDR available. Let$N(\lambda)$ be the number of observed p-values that are less than or equal to λ; let m be the number of tests; let $f = 1$ if the FINITE option is specified, and otherwise set $f = 0$ and denote the estimated proportion of true null hypotheses by

$$\hat{\pi}_0(\lambda) = \frac{m - N(\lambda) + f}{(1-\lambda)m}$$

The default estimate of FDR is

$$\widehat{\mathrm{FDR}}_\lambda(p) = \frac{\hat{\pi}_0(\lambda) p}{\max(N(p), 1)/m}$$

If you set $\lambda = 0$, then this is identical to the FDR adjustment.

The positive FDR is estimated by

$$\widehat{\mathrm{pFDR}}_\lambda(p) = \frac{\widehat{\mathrm{FDR}}_\lambda(p)}{1-(1-p)^m}$$

The finite-sample versions of these two estimators for independent null p-values are given by

$$\widehat{\mathrm{FDR}}^*_\lambda(p) = \begin{cases} \dfrac{\hat{\pi}^*_0(\lambda)p}{\max(N(p),1)/m} & \text{if } p \le \lambda \\ 1 & \text{if } p > \lambda \end{cases}$$

$$\widehat{\mathrm{pFDR}}^*_\lambda(p) = \frac{\widehat{\mathrm{FDR}}^*_\lambda(p)}{1-(1-p)^m}$$

Finally, the adjusted p-values are computed as

$$\tilde{p}_i = \hat{q}_\lambda(p_i) = \inf_{p \ge p_i} \mathrm{FDR}_\lambda(p) \quad i = 1, \ldots, m$$

This method can produce adjusted p-values that are smaller than the raw p-values. This means that, as defined, controlling the false discovery rate enables you to reject these tests at a level less than the observed p-value. However, by default, this reduction is prevented with an additional restriction: $\tilde{p}_i = \max\{\tilde{p}_i, p_i\}$.

Chapter 9

Power and Sample Size for Multiple Comparisons

Westfall discusses this topic in chapter 7. As he indicates the best way to insure that the power is correct is by properly designing your experiment and pre-determining which tests are going to be performed before the analysis begins. As Westfall indicates power is harder to define and control in multiple comparison tests as there are several issues (or definitions) that one needs to consider.

9.1 Definitions of Power

In a single test of hypothesis the power is defined as,

$$Power = \Upsilon(\theta) = \Pr[\text{reject } H_0 \mid H_0 \text{ is false}]$$

where θ represents the parameter of interest in the problem. For example if one is testing $H_0 : \mu_1 = \mu_2$ then the parameter is $\theta = \mu_1 - \mu_2$ where one might wish to determine the sample size such that the power is at least .8 when $\theta = 5$. In multiple comparisons there are other definitions of power which include;

- Complete power = Pr[reject all H_{0i} that are false].
- Minimal power = Pr[reject at least one H_{0i} that is false].
- Individual power = Pr[reject a particular H_{0i} that is false].
- Proportional power = average proportion of false H_{0i} that are rejected.

9.2 Examples Using Individual Power

Suppose one wished to use Tukey's method in a one-way ANOVA where one wishes to detect meaningful differences between $\theta = \mu_i - \mu_{i'}$. In order to compute the individual power one must specify the following:

- $d = \theta$ is the "meaningful difference".
- $\hat{\sigma}$ is an estimate of the within group standard deviation.
- g is the number of groups.
- n is the within group sample size.
- α is the desired FWE for the test.

Westfall Chapter 7 Example

The SAS file is westfall-ch7.sas

```
                Westfall Chapter 7
Method=TUKEY, Nominal FWE=0.05, nrep=1000, Seed=12345
       True means = (10, 5, 5, 0, 0), n=10, s=5

   Quantity              Estimate    ---95% CI----

   Complete Power         0.00700    (0.002,0.012)
   Minimal Power          0.97500    (0.965,0.985)
   Proportional Power     0.44263    (0.431,0.454)
   True FWE               0.01200    (0.005,0.019)
   Directional FWE        0.01200    (0.005,0.019)
```

SAS provide two procedures for computing power and/or sample sizes. The procedures are PROC POWER and PROC GLMPOWER. Both will be briefly discussed.

9.3 SAS POWER Procedure

Power and sample size analysis optimizes the resource usage and design of a study, improving chances of conclusive results with maximum efficiency. The POWER procedure performs prospective power and sample size analyses for a variety of goals, such as the following:

- determining the sample size required to get a significant result with adequate probability (power)
- characterizing the power of a study to detect a meaningful effect
- conducting what-if analyses to assess sensitivity of the power or required sample size to other factors

Here prospective indicates that the analysis pertains to planning for a future study. This is in contrast to retrospective power analysis for a past study, which is not supported by the procedure.
A variety of statistical analyses are covered:

- t tests, equivalence tests, and confidence intervals for means
- tests, equivalence tests, and confidence intervals for binomial proportions
- multiple regression
- tests of correlation and partial correlation
- one-way analysis of variance
- rank tests for comparing two survival curves
- logistic regression with binary response
- Wilcoxon-Mann-Whitney (rank-sum) test

9.4 SAS GLMPOWER Procedure

This section describes the approaches used in PROC GLMPOWER to compute power and sample size.

Contrasts in Fixed-Effect Univariate Models

The univariate linear model is

$$Y = X\beta + \epsilon$$

In PROC GLMPOWER, the model parameters β are not specified directly. Instead one specifies y^*, which represents either conjectured response means or typical response values for each design profile. The vector β is obtained from y^* as,

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

Note that, in general, there is not a 1 to 1 correspondence between y^* and $\hat{\beta}$, in that, different scenarios for y^* may lead to the same $\hat{\beta}$. Care needs to be taken, especially when the model contains interaction terms.

SAS parameterizes the design matrix X in three parts, $(\ddot{X}, w, N)$, where

- The $q \times p$ essence design matrix $\ddot{X}$ is the collection of unique rows of X. Its rows are sometimes referred to as "design profiles." Here, $q \leq N$ is defined simply as the number of unique rows of X.
- The $q \times 1$ weight vector w reveals the relative proportions of design profiles. Row i of $\ddot{X}$ is to be included in the design w_i times for every w_j times row j is included. The weights are assumed to be standardized (i.e., sum up to 1).
- The total sample size is N. This is the number of rows in X. If you gather N $w_i = n_i$ copies of the i^{th} row of $\ddot{X}$, for $i = 1, \ldots, q$, then you end up with X.

It is useful to express the the crossproduct matrix $X'X$ in terms of these three parts,

$$X'X = N\ddot{X}'W\ddot{X}$$

where $W = diag(w)$. This representation of $X'X$ allows one to describe the portion (N) depending on sample size and the portion $(\ddot{X}'W\ddot{X})$ depending on the design structure.

A general linear hypothesis is,

$$H_0 : L\beta = \theta_0$$

where L is a full row rank $r_L \times p$ contrast matrix, θ_0 and is the null value (usually, $\theta_0 = 0$).

The test statistic is

$$F = \frac{SS_H/r_L}{\hat{\sigma}^2} \sim F(r_L, df_e)$$

where

$$\begin{aligned} SS_H &= \frac{1}{N}\left[(L\hat{\beta} - \theta_0)'(L(X'X)^{-1}L')^{-1}(L\hat{\beta} - \theta_0)\right] \\ \hat{\beta} &= (X'X)^{-1}X'y \\ \hat{\sigma}^2 = MS_E &= (y - X\hat{\beta})'(y - X\hat{\beta})/df_e \end{aligned}$$

when the null hypothesis is true and $df_e = N - rank(X)$.

If $H_A : L\beta \neq \theta_0$ is true then $F \sim F(r_L, df_e, \lambda)$ is distributed as a non-central F with noncentrality

$$\lambda = N(L\beta - \theta_0)'(L(\ddot{X}'W\ddot{X})^{-1}L')^{-1}(L\beta - \theta_0)(\sigma)^{-2}$$

Muller and Peterson (1984) give the power of the test as

$$power = \Pr[F(r_L, df_e, \lambda) \geq F_{1-\alpha}(r_L, df_e)]$$

Sample size is computed by inverting the power equation. Refer to Muller et al. (1992) and O'Brien and Shieh (1992) for additional discussion.

9.5 Effect Size

Effect Size Measures for F Tests in GLM

A significant test in a linear model indicates that the effect of the term or contrast being tested might be real. The next thing you want to know is, How big is the effect? Various measures have been devised to give answers to this question that are comparable over different experimental designs. If you specify EFFECTSIZE option in the MODEL statement, then GLM adds three measures of effect size:

- the noncentrality parameter for the F test
- the proportion of total variation accounted for (also known as the semipartial correlation ratio or the squared semipartial correlation)
- the proportion of partial variation accounted for (also known as the full partial correlation ratio or the squared full partial correlation)

The adjectives "semipartial" and "full partial" might seem strange. They refer to how other effects are "partialed out" of the dependent variable and the effect being tested. For "semipartial" statistics, all other effects (excluding the dependent variable y) are partialed out of the effect in question. This measures the (adjusted) effect as a proportion of the total variation in the dependent variable. For "full partial" statistics, all other effects (including the dependent variable y) are partialed out of the effect in question. This measures the (adjusted) effect as a proportion of only the dependent variation remaining after partialing, or the partial variation.

Noncentrality Parameter

The noncentrality parameter is directly related to the true distribution of the F-statistic when the null hypothesis is false. The uniformly minimum variance unbiased estimate (UMVUE) for the noncentrality is

$$NC_{\text{UMVUE}} = \frac{df \times (df_e - 2) \times FValue}{df_e} - df$$

where $FValue$ is the observed value of the F-statistic for the test and df and df_e are the numerator and denominator degrees of freedom, respectively. An alternative biased estimate with smaller expected mean square error is

$$NC_{\text{minMSE}} = \frac{df \times (df_e - 4) \times FValue}{df_e} - \frac{df \times (df_e - 4)}{df_e - 2}$$

(see Perlman and Rasmussen (1976) cited in Johnson, Kotz, and Balakrishnan (1994)). A $p \times 100\%$ lower confidence bound for the noncentrality is given by the value of NC for which $probf(FValue, df, df_e, NC) = p$, where $probf()$ is the cumulative probability function for the non-central F-distribution. This result can be used to form a $(1 - \alpha) \times 100\%$ confidence interval for the noncentrality.

Partial Proportion of Variation

The partial proportion of variation accounted for by the effect being tested is easiest to define by its natural sample estimate,

$$\hat{\eta}^2_{partial} = \frac{SS}{SS + SS_E}$$

where SS_E is the error sum of squares. An alternative estimate that is asymptotically unbiased is

$$\omega^2_{partial} = \frac{SS - df \times MS_E}{SS + (N - df) \times MS_E}$$

where $MS_E = SS_E/df_e$ is the mean square for error and N is the number of observations. The true $\eta^2_{partial}$ is related to the true noncentrality parameter NC by the formula

$$\eta^2_{partial} = \frac{NC}{NC + N}$$

Proportion of Total Variation

The proportion of total variation is define by its sample estimate, known as the (semipartial) ω^2 statistic,

$$\hat{\eta}^2 = \frac{SS}{SS_{C.total}}$$

where $SS_{C.total}$ is the corrected total sum of squares, and SS is the sum of squares due to the effect being tested. As with $\hat{\eta}^2_{partial}$, $\hat{\eta}^2$ is actually a biased estimate of the true η^2 an alternative that is approximately unbiased is the

$$\omega^2 = \frac{SS - df \times MS_E}{SS_{C.total} + MS_E}$$

where $MS_E = SS_E/df_e$ is the sample mean square for error. Whereas $\eta^2_{partial}$ depends only on the noncentrality for its associated F-test, the presence of the total sum of squares in the previous formulas indicates that η^2 depends on the noncentralities for all effects in the model. An exact confidence interval is not available, but if you write the formula for $\hat{\eta}^2$ as

$$\hat{\eta}^2 = \frac{SS}{SS + (SS_{C.total} - SS)}$$

then a conservative confidence interval can be constructed as for $\eta^2_{partial}$, treating $SS_{C.total} - SS$ as SS_E and $N - df - 1$ as the df_e Smithson (2004). This confidence interval is conservative in the sense that it implies values of the true η^2 that are smaller than they should be.[1]

[1] When interpreting the actual values of effect size measures, the approximately unbiased ω^2 estimates are usually preferred for point estimates. Some authors have proposed certain ranges as indicating "small," "medium," and "large" effects Cohen (1988), but general benchmarks like this depend on the nature of the data and the typical signal-to-noise ratio; they should not be expected to apply across various disciplines. For example, while an ω^2 value of 10% might be viewed as "large" for psychometric data, it can be a relatively small effect for industrial experimentation. Whatever the standard, confidence intervals for true effect sizes typically span more than one category, indicating that in small experiments, it can be difficult to make firm statements about the size of effects.

Part IV

Likelihood Models

Chapter 10

Mixed Models

The material presented in this chapter represents a fundimental change in how one estimates parameters in the linear model. Previously, the methods given in PROC GLM use least squares estimates where the normality assumption is added in order to preform statistical inference. The methods given in PROC MIXED are based upon the likelihood function when the dependent variable y is normally distributed. SAS refers to these models as **Mixed Linear Models**.

Overview

A mixed linear model is a generalization of the standard linear model in that dependent variable y can be correlated and have heterogenous variance. This procedure allows one to model the means of the data (as in the standard linear model) in addition to the variances and covariances. The assumptions for the model are

- The data are normally distributed (Gaussian).
- The population means of the dependent variable y are linear functions of specified parameters (fixed effects).
- The variances and covariances of y are functions of a separate (distinct) set of parameters (random effects).

The complete probability distribution of the data is determined by the above assumptions. The fixed-effects parameters are associated with known explanatory variables (covariates), as in the standard linear model. These variables can be either qualitative (analysis of variance), quantitative (linear regression), or combined (analysis of covariance). The mixed linear model is distinguishable from the standard linear model by its ability to model variance/covariance structures.

10.1 Notation – Mixed Model

The standard linear model is

$$y = X\beta + e \tag{10.1}$$

where the focus is to parametrize the mean $\mu = E[y]$ as a function of fixed-effects parameters β. The unobserved errors e are assumed to be independent and identically distributed Gaussian random variables with mean 0 and homogeneous variance σ^2.

The mixed model is

$$y = X\beta + Z\gamma + \varepsilon \tag{10.2}$$

where γ is an unknown vector of random-effects parameters with specified design matrix Z, and ε is an unobserved random error vector whose elements are no longer required to be independent and/or homogeneous. That is, the marginal univariate mixed model satisfies

$$y_i = x_i'\beta + z_i'\gamma_i + \varepsilon_i, \tag{10.3}$$

where $\gamma_i \sim N(0, G)$, $\varepsilon_i \sim N(0, R_i)$, when $COV(\gamma_i, \varepsilon_i) = 0$ for all i. It follows that,

$$y_i \sim N(x_i'\beta, \Sigma_{y_i} = z_i G z_i' + R_i). \tag{10.4}$$

[Note: if $R_i = \sigma^2 I$ and $z_i' = 0$, then Equation (10.4) reduces the usual assumption for Equation (10.1).

10.1.1 Estimation of the Marginal Model

Let α denote the vector of all variance and covariance parameters found in Σ_{y_i}. Let $\theta = (\beta', \alpha')'$ be the s dimensional vector of all parameters in the marginal model for y_i. The **likelihood function** for the normal data is,

$$L_{mle}(\theta) = \prod_{i=1}^{n} \left\{ (2\pi)^{-n/2} \mid V_i(\alpha) \mid^{-1/2} \times \exp\left(-1/2(y_i - x_i'\beta)' V_i(\alpha)^{-1}(y_i - x_i'\beta)\right) \right\},$$

where $V_i(\alpha) = \Sigma_{y_i}$.

If one assumes that α is known (which it never is!), then the **maximum likelihood estimate (m.l.e.)** for β, conditioned on the value of α is

$$\widehat{\beta}(\alpha) = \left(\sum_{i=1}^{n} x_i' W_i x_i \right)^{-1} \sum_{i=1}^{n} x_i' W_i y_i,$$

where $W_i = V_i(\alpha)^{-1}$.

If α is unknown (which is always the case!) then one uses an estimate for α given by $\hat{\alpha}$ (this is the idea of the working covariance matrix) and $\widehat{W}_i = V_i(\hat{\alpha})^{-1}$.

10.1.2 Restricted Maximum Likelihood Estimation (REML)

Verbeke and Molenberghs introduce this idea by considering a familiar example. Suppose one has a random sample of size n where $y_i \sim N(\mu, \sigma^2)$. When μ is known then the m.l.e. for σ^2 is $\hat{\sigma}^2 = \sum_i (y_i - \mu)^2/n$, which is unbiased for σ^2. When μ is unknown and estimated by $\bar{y}$, the m.l.e. for σ^2 becomes $\hat{\sigma}^2 = \sum_i (y_i - \bar{y})^2/n$ which is biased downward from σ^2, since $E[\hat{\sigma}^2] = (1 - 1/n)\sigma^2$. Yet, one can easily obtain an unbiased estimate for σ^2 as $s^2 = \sum_i (y_i - \bar{y})^2/(n-1)$, so it follows that one should be able to find an unbiased estimate for σ^2 without having to estimate μ first.

Let $U = A'Y$, where Y is the vector of observations $Y' = (y_1, y_2, \ldots, y_n)$ and A is any $n \times (n-1)$ matrix of rank $n-1$ that is orthogonal to the vector $\mathbf{j_n}$.[1] It follows that $U \sim N(0, \sigma^2 A'A)$ and that the m.l.e. for σ^2 is $\hat{\sigma}^2 = Y'A(A'A)^{-1}A'Y/(n-1)$ which can be shown to equal s^2 for any A satisfying the above restriction. Hence, this estimate is called the **Restricted Maximum Likelihood Estimate (REML)** since it is restricted to the conditions placed on the matrix A.

[1] $A'\mathbf{j_n} = \mathbf{0}$

REML Estimation of σ^2

Suppose one has the usual linear model,

$$y = X\beta + e$$

where X is $n \times p$ and $e_i \sim N(0, \sigma^2)$. The m.l.e. for σ^2 is,

$$\hat{\sigma}^2_{mle} = (Y - X(X'X)^{-1}X'Y)'(Y - X(X'X)^{-1}X'Y)/n = Y'(I - H)Y/n,$$

which can easily be shown to be biased downward by a factor of $(n-p)/n$. As before define $U = A'Y$ where A is any $n \times (n-1)$ matrix of rank $n-1$ that is orthogonal to the vector $\mathbf{j_n}$[2]. From which one has,

$$\hat{\sigma}^2_{rmle} = (Y - X(X'X)^{-1}X'Y)'(Y - X(X'X)^{-1}X'Y)/(n-p) = Y'(I - H)Y/(n-p).$$

This follows since the first column of X is $\mathbf{j_n}$ and $(I - H)\mathbf{j_n} = \mathbf{0}$.

REML Estimation for the Linear Mixed Model

Consiider the mixed model

$$y = X\beta + Z\gamma + \varepsilon.$$

where γ is an unknown vector of random-effects parameters with known design matrix Z, and ε is an unknown random error vector (whose elements are no longer required to be independent and homogeneous).

Let $U = A'Y$, where A is any $n \times (n-p)$ matrix with columns that are orthogonal to the columns of X. Then is follows that $U \sim N(0, A'V(\alpha)A)$ which does not depend upon β. A natural choice for A is $I - H$, where $H = X(X'X)^{-1}X'$. In which case, $u_i = y_i - x_i'\hat{\beta} = r_i$ where r_i is the least squares residual for the fixed effects model. It can be shown that,

$$\log L_{mle}(\theta) = -1/2 \log \mid V \mid -1/2 r'V^{-1}r - n/2 \log(2\pi)$$

and

$$\log L_{reml}(\theta) = -1/2 \log \mid V \mid -1/2 r'V^{-1}r - 1/2 \log \mid X'V^{-1}X \mid -(n-p)/2 \log(2\pi)$$

10.1.3 Model Fitting Procedures

The EM algorithm can be used to estimate the random effect, however, the procedure is slow and convergence can be a problem whenever the maximum of the likelihood is on or near the boundary of the parameter space. The Newton-Raphson method is used for estimating the model parameters, since it is not as sensitive to the above issues.

10.1.4 Inference for the Fixed Effects

The estimate for the fixed effects vector, β is given by,

$$\widehat{\beta} = \left(\sum_{i=1}^{n} x_i' W_i x_i \right)^{-1} \sum_{i=1}^{n} x_i' W_i y_i,$$

in which the unknown vector, α of variance components is replaced by its ML or REML estimates. Under the marginal model (conditioned on α), $\widehat{\beta}(\alpha)$ has a normal distribution with mean vector β and covariance,

$$Var(\widehat{\beta}) = \left(\sum_{i=1}^{n} x_i' W_i x_i \right)^{-1}.$$

[2]Let $A = (I - H)$.

Approximate Wald's Test

Test the hypothesis $H_0 : L'\beta = 0$ with

$$\text{Wald} = (\widehat{\beta} - \beta)' \Big[L \left(\sum_{i=1}^{n} x_i' W_i x_i \right)^{-1} L' \Big]^{-1} (\widehat{\beta} - \beta).$$

Wald has a null asymptotic chi-square distribution with $rank(L)$ degrees of freedom. The variance of $\widehat{\beta}$ is incorrect (too small) since it does not account for the variability in estimating α. As a result, one uses the t or F distribution where the numerator degrees of freedom are based upon $rank(L)$ and the denominator degrees of freedom (or the df for the t-distribution) are based upon the data. For example, Satterthwaite procedure.

10.1.5 Likelihood Ratio Tests

Suppose that the null hypothesis is $\beta \in \Theta_0 \subset \Theta$ where Θ is the parameter space for β. Let L_0 denote the maximum value of the likelihood function when the null hypothesis is true ($\beta \in \Theta_0$) and L_1 denote the maximum value of the likelihood function maximum when ($\beta \in \Theta$), then

$$-2ln\lambda_N = -2ln\Big[\frac{L_0}{L_1}\Big]$$

has a asymptotic chi-square distribution with degrees of freedom equal to the difference in the dimension for Θ_0 and Θ. It should be mentioned that the asymptotic results hold for the ML estimation but not for REML.

10.2 Best Linear Unbiased Prediction (BLUP)

The main focus of the least squares design models has been to estimate of the fixed effects where the estimates of the random effects have been primarily used to determine the standard errors for the fixed effects. The mixed model methodology enables one to estimate (predict) specific random effects or linear functions of the random effects. The **Best Linear Unbiased Prediction (BLUP)** are the estimates of the variance components using the likelihood methodology.

McCulloch and Searle consider the simple example, $E[y_{ij} \mid \alpha_i] = \mu + \alpha_i$ where α_i represents the treatment given at a randomly selected medical center i. Suppose that one wants to gain information about a particular center (α_i). It can be shown that the "best" estimate for α_i is the conditional mean $E[\alpha_i \mid y]$. Estimators are called "best" in that they have smaller mean squared error than estimators based on the assumption that the random effects are fixed effects. Estimates that have this property are called *shrinkage estimators*. That is,

$$\begin{aligned} var(\alpha) &= var(E[\alpha \mid y]) + E[var(\alpha)] \\ &= var(\tilde{\alpha}) + \text{ a positive value} \end{aligned}$$

where $\tilde{\alpha} = E[\alpha \mid y]$ is the predictor. Thus,

$$var(\tilde{\alpha}) \leq var(\alpha).$$

10.2.1 Basic Concepts – BLUP

Assume that $y = X\beta + Z\gamma + e$, where β contains the fixed effects, γ contains the random components, and $\gamma \sim N(0, G)$. The **unconditional** mean and variance are

$$\begin{aligned} E[y] &= X\beta \\ var(y) &= ZGZ' + R \end{aligned}$$

and the **conditional** mean and variance are

$$\begin{aligned} E[y \mid \gamma] &= X\beta + Z\gamma \\ var(y \mid \gamma) &= R \end{aligned}$$

The unconditional mean is a population-wide average whereas the conditional mean is an average specific to an observed set of random effects (Littell).

Linear combination of the fixed effects are said to be estimable if $K'\beta = T'E[y] = T'X'\beta$ or $K = XT$ for some T. Estimable functions do not depend upon the random effects γ, however, there are cases when one wants to estimate linear combinations of both the fixed and random effects, $K'\beta + M'\gamma$. This function is said to be **predictable** if $K'\beta$ is estimable. Suppose that one has estimates $\hat{\beta}$ and $\hat{\gamma}$ then $K'\hat{\beta} + M'\hat{\gamma}$ is said to be the best unbiased predictor (BLUP) of $K'\beta + M'\gamma$.

Chapter 11

Loglinear Models for Contingency Tables

Previously, test of hypotheses with two-way contingency tables were used to determine whether or not two variables were statistically independent or associated. In this chapter the notion of two variables being independent or associated is considered from a modeling perspective. The text entitled, "Categorical Data Analysis Using the SAS System," by Stokes, Davis, and Kock provides an excellent resource for problems of this type.

11.1 Loglinear Models – Two-Way Tables

2×2 Tables

Suppose that one has the 2×2 table

X	Y=1	Y=2	Total
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

with cell probabilities given as

X	Y=1	Y=2	Total
1	π_{11}	π_{12}	π_{1+}
2	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

The analysis of primary interest with tables of this type is to investigate the independence of the two random variables where X and Y are independent implies that,

$$\frac{\pi_{11}}{\pi_{+1}} = \frac{\pi_{12}}{\pi_{+2}} = \pi_{1+}$$

and

$$\pi_{11} = \pi_{1+}\pi_{+1}.$$

From which it follows that

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad i, j = 1, 2.$$

The expected number of observations in the ij^{th} cell is

$$m_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j} = \mu\alpha_i\beta_j$$

where α_i represents the effect for X and β_j represents the effect for Y. Taking the log of this term gives

$$\begin{aligned} \log(m_{ij}) &= \log(\mu) + \log(\alpha_i) + \log(\beta_j) \\ &= \lambda + \lambda^X + \lambda^Y \end{aligned}$$

when X and Y are independent. This expression becomes

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

or

$$m_{ij} = \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY})$$

for $i, j = 1, 2$ when X and Y are not independent and λ^{XY} denotes the term for the multiplicative (association) term. This equation is called the **saturated loglinear model** for the 2×2 table. Since there are $1 + 2 + 2 + 4 = 9$ parameters in this model and only four observations (cell frequencies) it is necessary to define the following constraints on the model,

$$\sum_i \lambda_i^X = 0, \quad \sum_j \lambda_j^Y = 0, \text{ and } \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0.$$

The loglinear model expected cell counts can be written as,

X	Y=1	Y=2
1	$exp(\mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY})$	$exp(\mu + \lambda_1^X - \lambda_1^Y - \lambda_{11}^{XY})$
2	$exp(\mu - \lambda_1^X + \lambda_1^Y - \lambda_{11}^{XY})$	$exp(\mu - \lambda_1^X - \lambda_1^Y + \lambda_{11}^{XY})$

If X and Y are independent then the odds ratio, ψ, is

$$\psi = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1$$

and the log odds can be written as

$$\log \psi = \log \pi_{11} + \log \pi_{22} - \log \pi_{12} - \log \pi_{21} = 0.$$

The estimated log odds ratio, $\hat{\psi} = \frac{m_{11}m_{22}}{m_{12}m_{21}}$, can be written as

$$\log \hat{\psi} = \log m_{11} + \log m_{22} - \log m_{12} - \log m_{21} = 4\lambda_{11}^{XY}$$

In which case, the hypothesis of independence of X and Y is equivalent to testing $H_0 : \lambda_{11}^{XY} = 0$.

11.2 Three-Way Tables

Types of Independence

- The **saturated model** the loglinear form is

$$log\ \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.$$

- **Mutual Independence** X, Y, and Z are mutually independent when

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$$

for all i, j and k. Mutual independence implies that the expected frequencies $\{\mu_{ijk}\}$ have the loglinear form is

$$log\ \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z,$$

since

$$\lambda_{ij}^{XY} = \lambda_{ik}^{XZ} = \lambda_{jk}^{YZ} = \lambda_{ijk}^{XYZ} = 0.$$

- **Joint Independence** Variable Y is jointly independent of X and Z when

$$\pi_{ijk} = \pi_{i+k}\pi_{+j+}$$

for all i, j and k. The loglinear form is

$$log\ \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ},$$

since

$$\lambda_{ij}^{XY} = \lambda_{jk}^{YZ} = \lambda_{ijk}^{XYZ} = 0.$$

- **Conditional Independence** X and Y are conditionally independent, given Z when

$$\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}$$

for all i, j and k. This holds for every joint probabilities as

$$\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}.$$

The loglinear form is

$$log\ \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ},$$

since

$$\lambda_{ij}^{XY} = \lambda_{ijk}^{XYZ} = 0.$$

Chapter 12

Generalized Linear Models

12.1 Introduction

The Generalized Linear Model (GLM) extends the linear regression/anova model to the case where the response variable Y may have a non-normal distribution. The problem of interest is, model a function, $g(\mu)$, for $E(Y) = \mu$ as a linear function of explanatory variables X, given by $\eta(X)$. The GLM model consists of three components; 1) the random component associated with the response variable, Y, 2) the systematic component associated with the explanatory variables, X, and 3) a link function that specifies the function, $g(\mu)$.

12.1.1 Components of the GLM

1. **The random component** specifies the probability distribution for the response variable, Y where the p.d.f. for Y, $f_Y(\cdot)$ is said to be **exponential class, :-)**. That is, $f_Y(\cdot)$ can be written as

$$f(y;\theta) = a(\theta)b(y)exp[t(y)\ Q(\theta)], \tag{12.1}$$

or

$$f(y;\theta,\phi) = exp\left(\frac{[y\ \theta - b(\theta)]}{a(\phi)} + c(y,\phi)\right). \tag{12.2}$$

Cassella and Berger use the notation given in equation (12.1). McCullagh and Nelder (1989) use the preferred notation as given in equation (12.2). This equation is called the *exponential class dispersion family* where ϕ is the *dispersion parameter* and θ is the *natural parameter*. Note: the functions $a(\cdot)$ and $b(\cdot)$ are not the same in equation (12.1) and equation (12.2).

2. **The systematic component** is defined as,

$$\eta_i = x_i'\beta = \sum_j \beta_j x_{ij},$$

for $i = 1, 2, \ldots, N$ and x_{ij} is the i^{th} response for the j^{th} independent variable (covariate).

3. **The link function component** defines the realtionship between the above components as, $\eta_i = g(\mu_i)$, where $\mu_i = E(Y_i)$ and $g(\cdot)$ is the link function. The link functions $\eta_i = g(\mu_i)$ can be any monotonic, differentiable function. However, in practice, only a small set of link functions are actually utilized. In particular, link functions are chosen such that; the inverse link, $\mu_i = g^{-1}(\eta_i)$ is easily computed,

and g^{-1} maps $\eta_i = x_i'\beta \in \Re$ into a set of admissible values for μ_i. For example, if $\mu = \pi \in (0,1)$ then $g^{-1}(x_i'\beta) \in (0,1)$ where $x_i'\beta \in \Re$. In which case it follows that,

$$\eta_i = g(\mu_i) = \sum_j \beta_j \; x_{ij}$$

or

$$\mu_i = g^{-1}(\sum_j \beta_j \; x_{ij}).$$

If $g(\mu) = \mu$, the link is called the identity link and the GLM is the usual least squares linear model when the distribution of Y is Gaussian.

12.1.2 Examples – GLM Distributions

Normal or Gaussian

The pdf for the $Y \sim N(\mu, \sigma^2)$ is

$$f(y; \mu, \sigma^2) = \sqrt{2\pi} \; \sigma \; exp[-1/2[(y-\mu)/\sigma]^2].$$

This function can be written [using equation (12.2)] as

$$exp\{(y\mu - \mu^2/2)/\sigma^2 - 1/2[y^2/\sigma^2 + log(2\pi\sigma^2)]\},$$

where $\theta = \mu$, $b(\theta) = \theta^2/2$, $a(\phi) = \sigma^2$, and $c(y;\phi) = 1/2[y^2/\sigma^2 + log(2\pi\sigma^2)]$. Since $\theta = g(\mu) = \mu$ the canonical link is the identity link function.

Bernoulli Logit Model

Suppose that Y is binary with $\Pr[Y=1] = \pi$ and $\Pr[Y=0] = 1-\pi$. The probability mass function, pmf, for the Bernoulli is

$$p(y;\pi) = \pi^y(1-\pi)^{1-y}.$$

The pmf can be written [using equation (12.2)] as

$$p(y;\pi) = \pi^y(1-\pi)^{1-y} = exp[y \; log[\pi/(1-\pi)] + log(1-\pi)] = exp[y \; \theta - log(1+e^\theta)],$$

where $\theta = log[\pi/(1-\pi)]$, $a(\phi) = 1$ and $b(\theta) = log(1+e^\theta)$. The natural or canonical link is the log odds, $g(\pi) = log[\pi/(1-\pi)]$, and $g(\cdot)$ is the **Logit** function. Models of this type are called the **Logistic model**.

Poisson Loglinear model

Suppose that the response variable, Y, are counts. A potential model for Y is the Poisson model with pmf given by,

$$p(y;\mu) = e^{-\mu}\mu^y/y!.$$

The pmf can be written [using equation (12.2)] as

$$p(y;\mu) = exp[y \; log(\mu) - \mu - log(y!)],$$

where $\theta = \log(\mu)$, $a(\phi) = 1$, $b(\theta) = e^\theta = \mu$, and $c(y;\phi) = -log(y!)$. The natural or canonical link is $log(\mu)$. Models of this type are called the **Poisson Loglinear model**.

Assignment with Geometric

The Geometric has two forms given by;

1. Number of trials needed for the first success -

$$f_Y(y|\pi) = (1-\pi)^{y-1}\pi$$

for $y = 1, 2, \ldots,$ where $\pi = \Pr[`success']$.

2. Number of failures before the first success -

$$f_Z(z|\pi) = (1-\pi)^{z}\pi$$

for $z = 0, 1, 2, \ldots,$

Show that these Geometric densities are exponential class. State $b(\theta)$ and the canonical link.

12.1.3 Moments for the GLM

The following result is very useful for the GLM Models.

Theorem 12.1 *If $Y \sim f_Y(\cdot)$ is a member of the exponential family with pdf given by equation (12.2) then it can be shown that,*

$$E(U_i) = E(\partial \mathcal{L}_i/\partial\theta_i) = 0 \tag{12.3}$$

and

$$\begin{aligned} Var(U_i) &= E[U_i^2] \\ &= -E(\partial^2 \mathcal{L}_i/\partial\theta_i^2) \\ &= E[(\partial \mathcal{L}_i/\partial\theta_i)^2] \end{aligned} \tag{12.4}$$

where

$$\mathcal{L}_i = log(f(y_i;\theta_i,\phi)) = \{[y_i\ \theta_i - b(\theta_i)]/a(\phi) + c(y_i,\phi)\},$$

and

$$U_i = \partial \mathcal{L}_i/\partial\theta_i$$

is the score equation. $Var(U_i)$ is called the Fisher's Information.

Proof: Equation (12.3) holds, since

$$U = \partial\mathcal{L}/\partial\theta = \frac{1}{f(y_i;\theta,\phi)}\frac{\partial}{\partial\theta}[f(y_i;\theta,\phi)]$$

and

$$E(U) = \int \frac{\partial}{\partial\theta}[log(f(y;\theta,\phi))](f(y;\theta,\phi))dy = \int \frac{\partial}{\partial\theta}[f(y;\theta,\phi)]dy. \tag{12.5}$$

By assuming some regularity conditions that allows one to interchange differentiation and integration [which hold under exponential class assumptions], one has equation. Note: these regularity conditions are the same as those used when proving the Cramer-Rao lower bound for unbiased estimators of $\tau(\theta)$ as in Cassella and Berger.

$$\int \frac{\partial}{\partial\theta}[f(y;\theta,\phi)]dy = \frac{\partial}{\partial\theta}\int f(y;\theta,\phi)dy = \frac{\partial}{\partial\theta}[1] = 0$$

Equation (12.4) follows by taking the partial derivative of equation (12.5). That is,

$$\frac{\partial}{\partial\theta}E(U) = \frac{\partial}{\partial\theta}\int\frac{\partial}{\partial\theta}[log(f(y;\theta,\phi))]f(y;\theta,\phi)dy = \frac{\partial}{\partial\theta}\int\frac{\partial}{\partial\theta}[f(y;\theta,\phi)]dy = 0.$$

By assuming another regularity assumption where

$$\frac{\partial}{\partial\theta}\int\frac{\partial}{\partial\theta}[log(f(y;\theta,\phi))]f(y;\theta,\phi)dy = \int\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta}[log(f(y;\theta,\phi))]f(y;\theta,\phi)dy.$$

Equation (12.4) holds, since

$$\begin{aligned}\int\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta}[log(f(y;\theta,\phi))]f(y;\theta,\phi)dy &= \int\{\frac{\partial^2}{\partial\theta^2}[log(f(y;\theta,\phi))](f(y;\theta,\phi))dy\\ &+ \int\frac{\partial}{\partial\theta}[log(f(y;\theta,\phi))]\frac{\partial}{\partial\theta}[f(y;\theta,\phi)]\}dy\\ &= 0.\end{aligned}$$

or

$$\int\frac{\partial^2}{\partial\theta^2}[log(f(y;\theta,\phi))]f(y;\theta,\phi)dy + \int[\frac{\partial}{\partial\theta}[log(f(y;\theta,\phi))]]^2f(y;\theta,\phi)dy = 0.$$

or

$$-E[\frac{\partial^2}{\partial\theta^2}[log(f(y;\theta,\phi))]] = E[(\frac{\partial}{\partial\theta}[log(f(y;\theta,\phi))])^2].$$

Using the above properties where

$$\partial\mathcal{L}_i/\partial\theta = [y_i - b'(\theta_i)]/a(\phi)$$

and

$$\partial^2\mathcal{L}_i/\partial\theta^2 = -b''(\theta)/a(\phi),$$

when $\mathcal{L}_i = [y_i\ \theta - b(\theta)]/a(\phi) + c(y_i,\phi)$. Then

$$E(y_i) = \mu_i = b'(\theta_i) \tag{12.6}$$

and

$$Var(y_i) = b''(\theta_i)a(\phi). \tag{12.7}$$

12.1.4 Examples – Moments for GLM

Normal

When $Y_i \sim N(\mu_i,\sigma_i^2)$ we have, $\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2$, $a(\phi_i) = \sigma_i^2$, and $c(y;\phi_i) = 1/2[y_i^2/\sigma_i^2 + log(2\pi\sigma_i^2)]$. In which case,

$$E(y_i) = b'(\theta_i) = \theta_i = \mu_i$$

and

$$Var(y_i) = b''(\theta_i)a(\phi_i) = a(\phi_i) = \sigma_i^2.$$

Bernoulli

When $Y_i \sim \text{Binomial}(n_i, \pi_i)$, we have

$$p(y_i;\pi) = \binom{n_i}{y_i}\pi^{y_i}(1-\pi)^{n_i-y_i} = \exp\,[\frac{y_i\theta_i - n_i\ log[1+e^{\theta_i}]}{1} + \log\binom{n_i}{y_i}],$$

where $\theta_i = log[\pi_i/(1-\pi_i)]$, $b(\theta_i) = n_i \log[1+e^{\theta_i}]$, and $a(\phi) = 1$. In which case, it follows that

$$E(y_i) = b'(\theta_i) = n_i e^{\theta_i}/(1+e^{\theta_i}) = n_i\pi_i$$

and

$$\begin{aligned} Var(y_i) &= b''(\theta_i)a(\phi) \\ &= n_i e^{\theta_i}/(1+e^{\theta_i})^2 \\ &= n_i e^{\theta_i}/(1+e^{\theta_i})[1/(1+e^{\theta_i})] \\ &= n_i\pi_i(1-\pi_i), \end{aligned}$$

since $\pi_i = e^{\theta_i}/(1+e^{\theta_i})$ and $(1-\pi_i) = 1/(1+e^{\theta_i})$.

Poisson

When $Y_i \sim \text{Poisson}(\mu_i)$ we have, $\theta_i = \log(\mu_i)$, $a(\phi) = 1$, $b(\theta_i) = \exp(\theta_i)$, and $c(y_i;\phi) = -log(y_i!)$. In which case,

$$E(y_i) = \mu_i = b'(\theta_i) = \exp(\theta_i) = \mu_i$$

and

$$Var(y_i) = b''(\theta_i)a(\phi) = \exp(\theta_i) = \mu_i,$$

since $a(\phi) = 1$.

Assignment Geometric (continue)

Derive the mean and variance for each Geometric density (Y and Z).

12.2 Formal Structure for GLM

Myers, Montgomery, and Vining (MMV) provide a useful outline summarizing the structure for the GLM on page 161. Their summary is;

1. Observe $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$ independent random variables with means $\mu_1, \mu_2, \ldots, \mu_n$.
2. The random variable Y_i has a distribution as described by equation (12.2).
3. The systematic portion of the model involves $p-1$ regressors, $x_1, x_2, \ldots, x_{p-1}$.
4. The model is constructed about the **linear predictor** $\eta_i = x_i'\beta = \beta_0 + \sum_{j=0}^{p-1}\beta_j\ x_{ij}$.
5. The model is found through the use of a **link function**,

   $$\eta_i = g(\mu_i), \quad i = 1, 2, \ldots, n.$$

   where

   $$E(Y_i) = g^{-1}(\eta_i) = g^{-1}(x_i'\beta).$$

6. The link function is a monotone, differentiable function.

7. The variance σ_i^2 is a function of the mean μ_i. That is, $Var(Y_i) = \sigma_i^2 = v(\mu_i)$ for some function $v(\cdot)$.

There are many choices of the link function. If $\eta_i = \theta_i$, then the link function, $g(\cdot)$, is called the **canonical link**. The table contains the canonical links for some commonly used distributions.

Canonical Links

Distribution	Canonical Link
Normal	$\eta_i = \mu_i$ (identity link)
Binomial	$\eta_i = log[\frac{\pi_i}{(1-\pi_i)}]$ (logistic link)
Poisson	$\eta_i = log(\mu_i)$ (log link)
Exponential	$\eta_i = 1/\mu_i$ (reciprocal link)
Gamma	$\eta_i = 1/\mu_i$ (reciprocal link)

Other link functions include:

1. The **probit**
$$\eta_i = \Phi^{-1}[E(y_i)]$$
where Φ represents the cdf for the standard normal distribution.

2. The **complimentary log-log**
$$\eta_i = log\{log[1 - \mu_i]\}.$$

3. The **power family**
$$\eta_i = \mu_i^{\lambda}\ I_{\{\lambda \neq 0\}}(\lambda) + log[\mu_i]\ I_{\{0\}}(\lambda).$$

12.2.1 Statistical Inference For Categorical Data

The Likelihood equation is given by

$$L(\theta) = L(\theta \mid y_1 \ldots, y_n) = \prod_i f_Y(y_i; \theta).$$

The log likelihood function is given by,

$$\mathcal{L}(\theta) = log(L(\theta)) = \sum_i \mathcal{L}_i(\theta)$$

where $\mathcal{L}_i(\theta) = \log f_Y(y_i; \theta)$. This function can be used to determine an estimate of θ, given by $\hat{\theta}$, that maximizes the likelihood function. $\hat{\theta}$ is often (but not always) determined by finding solutions to the **score equation** $U(\theta)$

$$U(\theta) = \partial\mathcal{L}(\theta)/\partial\theta = 0.$$

When θ is p dimensional, then the score equation defines a system of p equations

$$\partial\mathcal{L}(\theta)/\partial\theta_j = 0,$$

for $j = 1, 2, \ldots, p$. The standard error, denoted by $SE(\hat{\theta})$, is the positive square root of the diagonal elements of the covariance matrix, $cov(\hat{\theta})$. The asymptotic covariance of $\hat{\theta}$ is the inverse of the **Fisher's information matrix**, $\Im(\theta)$, where the $(jk)^{th}$ element of $\Im(\theta)$ is

$$\Im(\theta) = (\Im_{jk}) = -E\left(\frac{\partial^2\mathcal{L}(\theta)}{\partial\theta_j\partial\theta_k}\right).$$

Bring Appendix A – Review of Likelihood Theory as found on blackboard to class as we will cover this in class

12.2.2 Example – MLE for the Binomial Distribution

The log likelihood function for the binomial distribution (ignoring $\binom{n}{y}$) is

$$\mathcal{L}(\pi) = log[\pi^y(1-\pi)^{n-y}] = y\ log(\pi) + (n-y)\ log(1-\pi),$$

The score equation is

$$U(\pi) = \partial\mathcal{L}(\pi)/\partial\pi = \frac{y}{\pi} - \frac{(n-y)}{(1-\pi)}.$$

In which case, $\hat{\pi}_{mle} = y/n$. Taking the expectation of $\partial^2\mathcal{L}(\pi)/\partial\pi^2$ gives

$$\Im(\pi) = -E[\partial^2\mathcal{L}(\pi)/\partial\pi^2] = E[y/\pi^2 + (n-y)/(1-\pi)^2] = n/[\pi(1-\pi)].$$

In which case, the asymptotic variance for $\hat{\pi}_{mle}$ is $\pi(1-\pi)/n$.

12.2.3 Tests of Hypothesis

Three commonly used methods for testing $H_0 : \theta = \theta_0$ are; Wald's test, the likelihood ratio test, and the score test.

Wald's Test

Wald's test statistic is

$$z = (\hat{\theta} - \theta_0)/SE(\hat{\theta}) \rightarrow N(0,1). \tag{12.8}$$

Likelihood Ratio Test

Let $L_0 = \sup L(\theta \in \Theta_0)$ and $L_1 = \sup L(\theta \in \Theta)$ then the likelihood ratio statistic is

$$-2log\Lambda = -2log(L_0/L_1) = -2(\mathcal{L}_0 - \mathcal{L}_1) \rightarrow \chi^2(df), \tag{12.9}$$

where $\mathcal{L}_0 = log(L_0)$, $\mathcal{L}_1 = log(L_1)$, and df is equal to the difference in the dimension of the parameter space between Θ and Θ_0.

Score Statistics

The score test statistic is

$$\left[U(\theta_0) \times \sqrt{\Im(\theta_0)}\right] \rightarrow N(0,1). \tag{12.10}$$

Note, this statistic depends solely upon the value θ_0. Hence, one need not compute the maximum likelihood estimate, $\hat{\theta}$ (as with Wald and the likelihood test). Wald's test is more powerful than the score test.

12.2.4 Constructing Confidence Intervals

It is more instructive to construct confidence intervals for the parameters than to test hypothesis about values of the parameter that we know are (or hope to be) wrong.

Wald's Intervals

The confidence interval for θ is given by

$$\hat{\theta} \pm z_{\alpha/2} \times \sigma_{\hat{\theta}}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile for the standard normal distribution. Although this was one of the first used methods for constructing confidence intervals its performance is very poor unless the sample size n is large. This is especially true when the parameter of interest if $p = \Pr[\text{“}success\text{”}]$ in the binomial distribution is close to zero or one.

Likelihood Ratio Method

The confidence interval based upon the likelihood ratio is the set of all values θ_0 such that

$$-2[\mathcal{L}(\theta_0) - \mathcal{L}(\hat{\theta})] < \chi^2_1(\alpha)$$

where $\chi^2_1(\alpha)$ is the $100(1-\alpha)$ percentile for the chi-square distribution with 1 df. Recall: $\chi^2_1(\alpha) = z^2_{\alpha/2}$.

Score Method

The confident interval based upon the score procedure is the set of all values θ_0 such that

$$U(\theta_0) \times \sqrt{\Im(\theta_0)} < z_{(\alpha/2)}.$$

Note: If $\hat{\theta}$ has a normal distribution then the intervals constructed using the above methods will produce similar results. However, if the distribution of $\hat{\theta}$ is not normal the results for the likelihood ratio method will differ greatly from Wald's method with small samples. The likelihood method is the preferred method when the sample size is small.

12.2.5 Example – Inference for the Binomial Distribution

Test of Hypothesis: $H_0 : \pi = \pi_0$

- Wald's Test statistics is,

$$z_W = \frac{\hat{\pi} - \pi_0}{SE(\hat{\pi})} = \frac{(\hat{\pi} - \pi_0)}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}.$$

- The score test statistics is,

$$z_S = U(\pi_0) \times \sqrt{\Im(\pi_0)} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

  Since, $U(\pi_0) = y/\pi_0 - (n-y)/(1-\pi_0)$, and $\Im(\pi_0) = n/(\pi_0(1-\pi_0))$.

- The likelihood function for the binomial is (ignoring $\binom{n}{y}$),

$$\mathcal{L}(\pi) = log[\pi^y(1-\pi)^{n-y}] = y\ log(\pi) + (n-y)\ log(1-\pi),$$

  from which we have,

$$\mathcal{L}_0 = \mathcal{L}(\pi_0) = y\ log(\pi_0) + (n-y)\ log(1-\pi_0),$$

  and

$$\mathcal{L}_1 = \mathcal{L}(\hat{\pi}) = y\ log(\hat{\pi}) + (n-y)\ log(1-\hat{\pi}).$$

In which case the Likelihood Ratio test statistics is,

$$\begin{aligned} -2(\mathcal{L}_0 - \mathcal{L}_1) &= 2\left(y\ log(\hat{\pi}/\pi_0) + (n-y)\ log((1-\hat{\pi})/(1-\pi_0))\right) \\ &= 2\left(y\ log\frac{y}{n\pi_0} + (n-y)log\frac{n-y}{n-n\pi_0}\right) \\ &= 2\left(y\ log\frac{\hat{\pi}}{\pi_0} + (n-y)log\frac{1-\hat{\pi}}{1-\pi_0}\right), \end{aligned} \tag{12.11}$$

where $\hat{\pi} = y/n$.

Confidence Intervals

- Wald's CI is,

$$\hat{\pi} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile for the standard normal distribution.

- The Likelihood Procedure. The endpoints to the intervals are the solutions to

$$2\left(y\ log\frac{\hat{\pi}}{\pi_0} + (n-y)log\frac{1-\hat{\pi}}{1-\pi_0}\right) = \chi^2_1(\alpha/2).$$

- The Score Procedure. The endpoints to the intervals are solutions to

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = z_{(\alpha/2)}.$$

12.2.6 Example – MLE for the Multinomial Distribution

The pdf for the multinomial distribution is proportional to,

$$\prod_j \pi_j^{n_i}$$

where $\pi_j \geq 0$ and $\sum_j \pi_j = 1$. The log likelihood function is

$$\mathcal{L}(\pi) = \sum_j n_j\ log(\pi_j).$$

Differentiating with respect to π_j, where $\pi_c = 1 - \sum_{j=1}^{c-1} \pi_j$ and $\partial\pi_c/\partial\pi_j = -1$, gives the score equation

$$\partial\mathcal{L}(\pi)/\partial\pi_j = n_j/\pi_j - n_c/\pi_c = 0.$$

The solution is

$$\hat{\pi}_j/\hat{\pi}_c = n_j/n_c.$$

Since $\sum_j \pi_j = 1$, it follows that

$$1 = \sum_j \hat{\pi}_j = \hat{\pi}_c\ n/n_c,$$

and

$$\hat{\pi}_j = n_j/n,$$

for $j = 1, 2, \ldots, c$.

12.2.7 Likelihood Ratio Test for the Multinomial Distribution

Test the hypothesis $H_0 : \pi_j = \pi_{j_0}$ for $j = 1, 2, \ldots, c-1$. The Likelihood Ratio test is

$$G^2 = -2\log\Lambda = 2\sum n_j \log(\frac{\hat{\pi}_j}{\pi_{j_0}}) - 2\sum n_j \log(n_j/n\pi_{j_0}) \rightarrow \chi^2(df = c-1)$$

where

$$\Lambda = \frac{\prod_j (\pi_{j_0})^{n_j}}{\prod_j (\hat{\pi}_j)^{n_j}} = \frac{\prod_j (\pi_{j_0})^{n_j}}{\prod_j (n_j/n)^{n_j}}.$$

12.3 Likelihood Equations for the GLM

In the previous section the likelihood methods for some distribution were discussed. In this section, the likelihood estimates for the GLM is presented. The likelihood function for N independent observations (we are not assuming identical distributions) is

$$\mathcal{L}(\beta) = \sum_i \mathcal{L}_i = \sum_i \frac{y_i\ \theta_i - b(\theta_i)}{a(\phi)} + \sum_i c(y_i, \phi).$$

The **score equations** are

$$U(\beta) = \partial\mathcal{L}(\beta)/\partial\beta = 0$$

or

$$U_j = \sum_i \partial\mathcal{L}_i/\partial\beta_j = 0,$$

for $j = 0, 1, \ldots, p$. In order to perform this differentiation one makes use of the chain rule

$$\frac{\partial\mathcal{L}_i}{\partial\beta_j} = \frac{\partial\mathcal{L}_i}{\partial\theta_i}\ \frac{\partial\theta_i}{\partial\mu_i}\ \frac{\partial\mu_i}{\partial\eta_i}\ \frac{\partial\eta_i}{\partial\beta_j}$$

where $\partial\mathcal{L}_i/\partial\theta_i = [y_i - b'(\theta_i)]/a(\phi)$, $\mu_i = b'(\theta_i)$, $var(Y_i) = b''(\theta_i)a(\phi)$, and $\eta_i = \sum_j \beta_j x_{ij}$. it follows that,

$$\begin{aligned} \partial\mathcal{L}_i/\partial\theta_i &= (y_i - \mu_i)/a(\phi), \\ \partial\mu_i/\partial\theta_i &= b''(\theta_i) = var(Y_i)/a(\phi), \\ \partial\eta_i/\partial\beta_j &= x_{ij}. \end{aligned}$$

In which case, the score equations are

$$U_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i)\ x_{ij}}{var(Y_i)}\ (\frac{\partial\mu_i}{\partial\eta_i}) = 0, \quad j = 0, 1, \ldots, p-1. \tag{12.12}$$

Equation (12.12) is called **Fisher's score function**. These equations are nonlinear and must be solved numerically using iterative methods. Note: the term $(\frac{\partial\mu_i}{\partial\eta_i})$ is dependent upon the link function $g(\cdot)$.

12.3.1 Newton-Raphson Solution

This method requires the computation of a $p \times p$ Hessian matrix of second derivatives $H = (h_{uv})$, where

$$\begin{aligned} h_{uv} &= \frac{\partial^2 \mathcal{L}_i}{\partial \beta_u \partial \beta_v} \\ &= -\sum_{i=1}^{N} \frac{x_{iu} x_{iv}}{a(\phi)} \big[\frac{1}{v(\mu_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2 - (\mu_i - y_i) \big\{ \frac{1}{v(\mu_i)^2} (\frac{\partial \mu_i}{\partial \eta_i})^2 \frac{\partial v(\mu_i)}{\partial \mu_i} - \frac{1}{v(\mu_i)} (\frac{\partial^2 \mu_i}{\partial \eta_i^2}) \big\} \big]. \end{aligned}$$

The m^{th} approximation of $\hat{\beta}$ is given by

$$b^{(m)} = b^{(m-1)} - H_{(m-1)}^{-1} \, U_{(m-1)},$$

where $H_{(m-1)} = H_{\beta = b_{(m-1)}}$ and $U_{(m-1)}$ is the vector of first derivatives of $\mathcal{L}_i$ evaluated at $\beta = b_{(m-1)}$ and $v(\mu_i) = var(Y_i) a(\phi)$.

12.3.2 Fisher's Score Function Solution

This method is known as **Integrated Reweighed Least Squares (IRLS)** and it makes use of Fisher's Information Criteria. Fisher's Information, $\Im$, is the negative of the expected value of the Hessian matrix given by,

$$\Im = -E[H] = (\Im_{uv}).$$

A numerical solution for the maximum likelihood estimates is

$$b^{(m)} = b^{(m-1)} + \Im_{(m-1)}^{-1} \, U_{(m-1)},$$

where $\Im_{(m-1)} = \Im_{\beta = b_{(m-1)}}$ and $U_{(m-1)}$ is the vector of first derivatives of $\mathcal{L}_i$ evaluated at $\beta = b_{(m-1)}$.
The elements of $\Im$ are

$$\begin{aligned} \Im_{uv} &= E[U_u U_v] \\ &= E[\frac{\partial \mathcal{L}_i}{\partial \beta_u} \frac{\partial \mathcal{L}_i}{\partial \beta_v}] \\ &= E\big[\sum_{i=1}^{N} \frac{(y_i - \mu_i) \; x_{iu}}{var(Y_i)} \; \frac{\partial \mu_i}{\partial \eta_i} \sum_{i=1}^{N} \frac{(y_i - \mu_i) \; x_{iv}}{var(Y_i)} \; \frac{\partial \mu_i}{\partial \eta_i} \big] \\ &= \sum_{i=1}^{N} \frac{E[(y_i - \mu_i)]^2 \; x_{iu} \; x_{iv}}{[var(Y_i)]^2} \; (\frac{\partial \mu_i}{\partial \eta_i})^2 \\ &= \sum_{i=1}^{N} \frac{x_{iu} x_{iv}}{Var(Y_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2. \end{aligned}$$

It follows that,

$$\Im = X' W X,$$

where W is a diagonal matrix with elements,

$$w_i = \frac{1}{Var(Y_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2 = \frac{1}{v(\mu_i) a(\phi)} (\frac{\partial \mu_i}{\partial \eta_i})^2. \tag{12.13}$$

One can now define the iterative process as

$$X'WXb_{(m)} = X'WXb_{(m-1)} + U_{(m-1)}.$$

Let $\nu = (\nu_i) = (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$, in which case one has

$$X'WXb_{(m)} = X'WXb_{(m-1)} + X'W\nu_{(m-1)} = X'Wz$$

where the i^{th} element of z is

$$z_i = x_i' b_{(m-1)} + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}.$$

If $X'WX$ has rank p (which it should have), then

$$b_{(m-1)} = (X'WX)^{-1}X'Wz$$

is the solution to the weighted least squares model. However, in this case both z and W depend upon the solution b. Thus, prompting use of the iterative least squares procedure.

Maximum Likelihood with a Canonical Link

The complexity of using either computational method described above is greatly reduced when $g(\cdot)$ is the canonical link function. That is, $\eta_i = \theta_i = g(\mu_i) = x_i' \vec{\beta} = \sum_i \beta_i \; x_{ij}$ and

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i).$$

The score equations are

$$\begin{aligned} U_j &= \sum_{i=1}^{N} \frac{(y_i - \mu_i)\; x_{ij}}{var(Y_i)} \; \frac{\partial \mu_i}{\partial \eta_i} \\ &= \sum_{i=1}^{N} \frac{(y_i - \mu_i) x_{ij}}{a(\phi)} \\ &= 0 \end{aligned}$$

for $j = 0, 1, \ldots, p-1$, and $b''(\theta_i) = var(Y_i)/a(\phi)$.
Note: if $a(\phi)$ is constant for each y_i, then score equations are

$$\sum_i x_{ij} \; y_i = \sum_i x_{ij} \; \mu_i. \tag{12.14}$$

Taking the partial derivative of $\frac{\partial \mathcal{L}_i}{\partial \beta_j}$ when $g(\cdot)$ the canonical link results in

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{N} \frac{x_{ij}}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \beta_k} \right).$$

Since this expression does not depend the variables y_i, one has

$$\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_j \partial \beta_k} = E\left[\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_j \partial \beta_k} \right]$$

or

$$H = -\Im.$$

in which case, the Newton-Raphson and Fisher's scoring algorithms provide identical solutions when the link function is the canonical link.

12.3.3 Example – Comparison of the Two Methods with the Binomial

Newton-Raphson Method

Let $y \sim Bin(n, \pi)$ then the likelihood equation is,

$$\mathcal{L}(\pi) = y \log \pi + (n - y) \log(1 - \pi)$$

from which

$$U = \frac{(y - n\pi)}{\pi(1 - \pi)}, \quad H = -\left[\frac{y}{\pi^2} + \frac{(n - y)}{(1 - \pi)^2}\right].$$

The iterative solution is

$$\pi^{(t+1)} = \pi^{(t)} + \left[\frac{y}{(\pi^{(t)})^2} + \frac{(n - y)}{(1 - \pi^{(t)})^2}\right]^{-1} \frac{(y - n\pi^{(t)})}{\pi^{(t)}(1 - \pi^{(t)})}.$$

Fisher's Scoring Method

$$\Im = -\left[\frac{n}{\pi(1 - \pi)}\right].$$

The iterative solution is

$$\begin{aligned} \pi^{(t+1)} &= \pi^{(t)} + \left[\frac{n}{\pi^{(t)}(1 - \pi^{(t)})}\right]^{-1} \frac{(y - n\pi^{(t)})}{\pi^{(t)}(1 - \pi^{(t)})}. \\ &= \pi^{(t)} + \frac{(y - n\pi^{(t)})}{n} \\ &= \frac{y}{n} \\ &= \hat{\pi}. \end{aligned}$$

This method converges to $\hat{\pi} = y/n$ in a single iteration.

12.3.4 Asymptotic Properties for the MLE with the GLM

If the model assumptions, including the link function are correct, then the solution b is asymptotically unbiased, $E(b) \to \beta$ and

$$Var(b) \to \Im(b)^{-1} = [X'VX]^{-1}[a(\phi)]^2$$

where the information matrix $\Im(b)$ is

$$\Im(b) = var[a(\phi)^{-1}(X'(y - \mu))] = \frac{X'VX}{(a(\phi))^2}$$

for $V = diag(\sigma_i^2)$, σ_i^2 is a function of $E(y_i) = \mu_i$, and the link is canonical. The asymptotic covariance matrix is

$$Var(b) = \Im(b)^{-1} = [X'\Delta V \Delta X]^{-1}[a(\phi)]^2$$

when the link is non-canonical where $\Delta = diag(\delta_1, \delta_2, \ldots, \delta_n)$ is a diagonal matrix with entries

$$\delta_i = \frac{\partial \theta_i}{\partial \mu_i}.$$

12.3.5 Examples - Asymptotic SE with Canonical Links

Normal

Let $y_i \sim N(\mu_i, \sigma^2)$ then

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

Binomial

Let $y_i \sim Bin(\pi_i, n_i)$ then

$$var(b) = (X'VX)^{-1}$$

where

$$\sigma_i^2 = n_i\pi_i(1-\pi_i) = \frac{n_i e^{x_i'\beta}}{(1+e^{x_i'\beta})^2}$$

since

$$\pi_i = \frac{e^{x_i'\beta}}{(1+e^{x_i'\beta})} = \frac{1}{(1+e^{-x_i'\beta})}$$

Poisson

Let $y_i \sim Poisson(\lambda_i)$ then

$$var(b) = (X'VX)^{-1}$$

where $\lambda_i = \sigma_i^2 = e^{x_i'\beta}$.

12.4 Inference for GLM

The Wald, score, and likelihood-ratio methods can be used with the GLM whenever the sample size is sufficiently large. For example, Wald's test for $H_0 : \beta_j = 0$ is

$$z = b_j/se(b_j) \sim N(0,1)$$

when the sample size is sufficiently large.

A likelihood-ratio method is based upon computing the **Deviance**. The Deviance is defined by considering the *saturated model* where each observation y_i is used as an estimate of $\mu_i = E(y_i)$. This method provides a perfect (over) fit for the model using the observed data. Let $\tilde{\mu}_i = y_i$ denote the estimate for μ_i in the saturated model. Let $L(\tilde{\mu}; y)$ denote the value of the likelihood when using the saturated model. Now suppose that $\hat{\mu}_i$ denotes the maximum likelihood estimate of μ_i. Let $L(\hat{\mu}; y)$ denote the value of the likelihood evaluated at the maximum likelihood estimates. The Deviance is

$$\begin{aligned} D(y;\hat{\mu})/\phi &= -2[\mathcal{L}(\hat{\mu};y) - \mathcal{L}(\tilde{\mu}=y;y)] \\ &= 2\left(\sum_i [y_i\tilde{\mu}_i - b(\tilde{\mu}_i)] - [y_i\hat{\mu}_i - b(\hat{\mu}_i)]\right)/a(\phi). \end{aligned} \tag{12.15}$$

where $\mathcal{L}(\cdot; y) = \log L(\cdot; y)$. $a(\phi)$ is often a constant multiple of ϕ/w_i, in which case, the *scaled deviance* is

$$D(y;\hat{\mu})/\phi = 2\sum_i w_i[y_i(\tilde{\mu}_i - \hat{\mu}_i) - b(\tilde{\mu}_i) + b(\hat{\mu}_i)]/\phi. \tag{12.16}$$

The larger the scaled deviance the poorer the fit.

12.4.1 Examples - Deviance

Normal

Let $y_i \sim N(\mu_i, \sigma_i^2)$, then $\theta_i = \mu_i$, $b(\theta_i) = \theta_i^2/2$, $a(\phi) = \sigma_i^2$, and $c(y;\phi) = 1/2[y^2/\sigma_i^2 + log(2\pi\sigma_i^2)]$. In which case,

$$D(y;\hat{\mu}) = \sum_i (y_i - \hat{\mu}_i)^2/\sigma_i^2.$$

Poisson

Let $y_i \sim Poisson(\mu_i)$ then $\theta_i = \log \mu_i$ and $b(\hat{\theta}_i) = \exp(\hat{\theta}_i) = \hat{\mu}_i$. Let $\tilde{\theta}_i = \log y_i$ and $b(\tilde{\theta}_i) = \exp(\tilde{\theta}_i) = y_i$. In which case,

$$D(y;\hat{\theta}) = 2\sum_i [y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i].$$

Residuals for GLM

Two types of residuals are commonly used with the GLM.

Deviance residual

The deviance residuals are defined as

$$\sqrt{d_i} \times sgn(y_i - \hat{\mu}_i), \tag{12.17}$$

where

$$d_i = 2\ w_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)].$$

Pearson residual

The pearson residuals are defined as

$$e_i = (y_i - \hat{\mu}_i)/\sqrt{\widehat{var(Y_i)}}, \tag{12.18}$$

where $\widehat{var(Y_i)}$ is dependent upon the distribution.

Chapter 13

Logistic Models

13.1 GLM for Binary Data

Let Y be a binary response variable where $\Pr[Y = 1 \mid \mathbf{x}] = \pi(\mathbf{x})$ and $\Pr[Y = 0 \mid \mathbf{x}] = 1 - \pi(\mathbf{x})$ with covariates $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. There are several potential approaches to this modeling problem. One could use the ordinary least squares approach[1], called the **Linear Probability model**, given as,

$$\pi(\mathbf{x}) = \alpha + \beta'\mathbf{x}.$$

This model has a structural defect since $\pi(x)$ is not restricted to the interval $[0, 1]$ for all x. A better model is the **Logistic Regression Model**[2] given as,

$$y = log\left[\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right] = (\alpha + \beta'\mathbf{x}),$$

where y is the log odds and $\pi(\mathbf{x})$ is the probability of the event of interest for the covariate $\mathbf{x}$. It follows that,

$$\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = exp(\alpha + \beta'\mathbf{x}),$$

and

$$\pi(\mathbf{x}) = \frac{exp(\alpha + \beta'\mathbf{x})}{1 + exp(\alpha + \beta'\mathbf{x})}.$$

13.1.1 Logistic Regression with Categorical Predictors

SDK considered using the Logistic model with categorical predictors. The SAS file is

`coronary_example.sas`

Coronary Example Output

The model for the four discrete events is given by

[1] GLM with normal data and the identity link function.
[2] GLM with binary data and the canonical logit link.

Sex	ECG	$\Pr[CA\ Disease] = \theta_{hi}$	Odds of CA Disease
Females	< 0.1	$e^{\alpha}/(1+e^{\alpha})$	e^{α}
Females	≥ 0.1	$e^{\alpha+\beta_2}/(1+e^{\alpha+\beta_2})$	$e^{\alpha+\beta_2}$
Males	< 0.1	$e^{\alpha+\beta_1}/(1+e^{\alpha+\beta_1})$	$e^{\alpha+\beta_1}$
Males	≥ 0.1	$e^{\alpha+\beta_1+\beta_2}/(1+e^{\alpha+\beta_1+\beta_2})$	$e^{\alpha+\beta_1+\beta_2}$

Parameter	Estimate	SE	Interpretation
α	-1.17	0.485	log odds of coronary disease for females with ECG < 0.1
β_1	1.28	0.498	increment to log odds for males
β_2	1.05	0.498	increment to log odds for high ECG

Sex	ECG	Logit	Odds of CA Disease
Female	< 0.1	$\hat{\alpha} = -1.17$	$e^{\hat{\alpha}} = e^{-1.17} = 0.3089$
Female	≥ 0.1	$\hat{\alpha} + \hat{\beta}_2 = -0.12$	$e^{\hat{\alpha}+\hat{\beta}_2} = e^{-0.12} = 0.8867$
Male	< 0.1	$\hat{\alpha} + \hat{\beta}_1 = 0.10$	$e^{\hat{\alpha}+\hat{\beta}_1} = e^{0.10} = 1.11$
Male	≥ 0.1	$\hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 = 1.157$	$e^{\hat{\alpha}+\hat{\beta}_1+\hat{\beta}_2} = e^{1.157} = 3.18$

13.2 Model Selection in GLM

Agresti indicated that one can perform model selection with GLM as with the usual multiple regression models but warns that these procedures need to be used with caution as the terms retained by the selection procedures are highly influenced by sample size. He suggests using both forward and backward selection (hoping that both procedures end up with the same models). Consider the following example. The file is

`coronary_example_2.pdf`

13.3 Alternative Models

Alternative models for logistic regression are

$$\pi(\mathbf{x}) = \Phi(\alpha + \beta\mathbf{x})$$

where Φ is a continuous cdf[3]. These models arise from the tolerance distribution found in dose-survival models. For example, suppose that an insect is killed ($Y = 1$) if the dosage (amount of pesticide) $x > T$ and an insect is not killed ($Y = 0$) when $x \leq T$. Let

$$\pi(\mathbf{x}) = \Pr[Y = 1 \mid \mathbf{X} = \mathbf{x}] = F_T(\mathbf{x}) = \Pr[T \leq \mathbf{x}].$$

One then selects the appropriate binary distribution that has the same shape as the tolerance distribution, F. Let Ψ denote the standard cdf of the family containing F, so that

$$\pi(\mathbf{x}) = F_T(\mathbf{x}) = \Psi[(\mathbf{x} - \mu)/\sigma]$$

or

$$\Psi^{-1}[\pi(\mathbf{x})] = \alpha + \beta\mathbf{x}.$$

[3]GLM with binary data and the non-canonical probit link function.

13.3.1 Likelihood for the Alternative Models

Recall, from equation(12.12)

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)\ x_{ij}}{var(Y_i)}\ \frac{\partial \mu_i}{\partial \eta_i} = 0$$

for $j = 0, 1, \ldots, p$ that can be written as

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)} \psi_i \left(\sum_j \beta_j x_{ij} \right) = 0,$$

since

$$\frac{\partial \mu_i}{\partial \eta_i} = \psi(\eta_i) = \psi \left(\sum_j \beta_j x_{ij} \right)$$

when $\psi(\eta_i) = \partial \Psi(\eta_i)/\partial \eta_i$.

13.3.2 Probit Model

Historically, toxicological experiments often measure dosage as the log concentration where the tolerance distribution for the dosage is assumed to be approximately $N(\mu, \sigma^2)$ for unknown μ and σ^2. In which case,

$$\pi(x) = \Phi(\alpha + \beta\ x)$$

where Φ is the standard normal cdf, $\alpha = -\mu/\sigma$ and $\beta = 1/\sigma$. This model is called the *Probit model*.

13.3.3 Complementary Log-Log Model

The complementary log-log model provides an alternative model to the logit and probit that is asymmetric about 0.5 where

$$\pi(x) = 1 - \exp[-\exp(\alpha + \beta x)],$$

and

$$\log[-\log(1 - \pi(x))] = \alpha + \beta x.$$

Let x_1 and x_2 denote two values of the covariate, then

$$\log[-\log(1 - \pi(x_2))] - \log[-\log(1 - \pi(x_1))] = \beta(x_2 - x_1),$$

or

$$\frac{\log[1 - \pi(x_2)]}{\log[1 - \pi(x_1)]} = \exp[\beta(x_2 - x_1)].$$

In which case, one has

$$1 - \pi(x_2) = [1 - \pi(x_1)]^{\exp[\beta(x_2 - x_1)]}.$$

Example – Challenger Data

Littell, Stroup, and Freund (2002) in SAS for Linear Models discuss the NASA Challenger data for which they fit a logistic regression model using both PROC LOGISTIC and PROC GENMOD. These data were the subject of an investigation following the 1986 Challenger space shuttle disaster. The focus of the investigation concerned the suspected association between O-ring failure and low air temperature at launch. O-ring failure is called thermal distress (TD denotes the number of launches with thermal distress at a specified temperature X).

A video for the analysis of this data is on canvas.

13.4 Logit Models for Multinomial Responses

13.4.1 Baseline-Category Logit Models

Suppose that Y has one of C categories. Let $\pi_j(x) = \Pr[Y = j \mid x)$, where $\sum_j \pi_j(x) = 1$. The logit baseline model assumes that all the comparisons are made with a baseline or reference category (last category in this notation) as,

$$log\left(\frac{\pi_j}{\pi_C}\right) = \alpha_j + \beta_j' x,$$

for $j = 1, 2, \ldots, C-1$. One can compute the logit for any pair of categories as

$$\log\left(\frac{\pi_a}{\pi_b}\right) = log\left(\frac{\pi_a/\pi_C}{\pi_b/\pi_C}\right) = log\left(\frac{\pi_a}{\pi_C}\right) - log\left(\frac{\pi_b}{\pi_C}\right) = (\alpha_a - \alpha_b) + (\beta_a' - \beta_b')x.$$

where class C is the reference classification. Note one could designate any category as the reference class.

Example – Food for Alligators

Agresti discusses this example in detail on pages 268-272. The log odds of selecting invertebrate (food=2) instead of fish (food=1 is the reference category) is

$$log(\hat{\pi}_I/\hat{\pi}_F) = -1.5490 + 1.4582 - 1.6583 + 0.9372 + 1.1220.$$

One can estimate log odds for pairs of non-baseline items, such as, invertebrate to other as

$$log(\hat{\pi}_I/\hat{\pi}_O) = log(\hat{\pi}_I/\hat{\pi}_F) - log(\hat{\pi}_O/\hat{\pi}_F).$$

13.4.2 Estimating Response Functions

The response functions for $\pi_j(x)$ is

$$\pi_j(x) = \frac{\exp(\alpha_j + \beta_j' x)}{1 + \sum_{h=1}^{C-1} \exp(\alpha_h + \beta_h' x)}$$

since $\alpha_C = 0$ and $\beta_C = 0$.
As an example, the estimated probability that a large alligator in Lake Hancock (lake=1) has invertebrates as food is

$$\hat{\pi}_I = \frac{e^{-1.55-1.66}}{1 + e^{-1.55-1.66} + e^{-3.31+1.24} + e^{-2.09+070} + e^{-1.90+0.83}} = 0.023.$$

13.5 Ordinal Responses

Suppose that the response variables are ordinal. There are several methods for analyzing data of these type.

13.5.1 Cumulative Logit Models

Define

$$\Pr[Y \leq j \mid x] = \sum_{i=1}^{j} \pi_i(x),$$

The cumulative logit is defined as

$$\text{logit}\left(\Pr[Y \le j \mid x]\right) = \log\Big[\frac{\Pr[Y \le j \mid x]}{1 - \Pr[Y \le j \mid x]}\Big] = \log\Big[\frac{\sum_{i=1}^{j}\pi_i(x)}{\sum_{i=j+1}^{C}\pi_i(x)}\Big],$$

for $j = 1, 2, \ldots, C-1$.
For a given j, the above is a binary logit model where the first group consists of the first j categories and the second group consists of the remaining categories. This method will provide possibly $C-1$ models. A better model would be when all $C-1$ cumulative logit models can be described as a single model.

13.5.2 Proportional Odds Model

Define

$$\textit{logit}\left(\Pr[Y \le j \mid x]\right) = \alpha_j + \beta' x, \quad j = 1, 2, \ldots, C-1. \tag{13.1}$$

Each cumulative has its own intercept but uses a common β. This assumption leads to,

$$\Pr[Y \le k \mid X = x] = \Pr[Y \le j \mid X = x + (\alpha_k - \alpha_j)/\beta],$$

when using the linear logit model. In which case,

$$\begin{aligned} \text{logit}\left(\Pr[Y \le j \mid x_1]\right) - \text{logit}\left(\Pr[Y \le j \mid x_2]\right) &= & (13.2)\\ \log\Big[\frac{\Pr[Y \le j \mid x_1]/\Pr[Y > j \mid x_1]}{\Pr[Y \le j \mid x_2]/\Pr[Y > j \mid x_2]}\Big] &= \beta'(x_1 - x_2). & (13.3)\end{aligned}$$

The odds of having a response $\le j$ at $X = x_1$ is $\exp[\beta'(x_1 - x_2)]$ times the odds of having this response at $X = x_2$. The log cumulative odds ratio is proportional to the distance between x_1 and x_2. The same proportionality constant applies to each logit.

13.6 Conditional Logistic Models

This material is given in The maximum likelihood procedure is optimal when the sample sizes are large, however, other procedures are better when the sample size is small. One such method is called *conditional maximum likelihood*.

13.6.1 Conditional Likelihood

Let Y_i denote the binary response for the i^{th} subject, where

$$P[Y_i = y_i] = \frac{\exp[y_i(\alpha + \sum_{j=1}^{p}\beta_j x_{ij})]}{1 + \exp(\alpha + \sum_{j=1}^{p}\beta_j x_{ij})}.$$

When $y_i = 1$ we have the usual logistic model where

$$\pi(x_{ij}) = P[Y_i = 1] = \frac{\exp[(\alpha + \sum_{j=1}^{p}\beta_j x_{ij})]}{1 + \exp(\alpha + \sum_{j=1}^{p}\beta_j x_{ij})}.$$

In which case the likelihood function is

$$P[Y_1 = y_1, \ldots, Y_N = y_N] = \frac{\exp[(\sum_{i=1}^{N} y_i)\alpha + \sum_{j=1}^{p}(\sum_{i=1}^{N} y_i x_{ij})\beta_j]}{\prod_{i=1}^{N}[1 + \exp(\alpha + \sum_{j=1}^{p}\beta_j x_{ij})]}.$$

From this expression it follows that the sufficient statistic for β_j is $\sum_i y_i x_{ij}$ and the sufficient statistic for α is $\sum_i y_i$, the total number of successes. Tis dependency upon the parameter α can be eliminated by conditioning with respect to the sufficient statistics for α. That is, let

$$S(t) = \{(y_1^*, \ldots, y_N^*) : \sum_i y_i^* = t\}$$

then the conditional likelihood becomes

$$\begin{aligned} P[Y_1 = y_1, \ldots, Y_N = y_N \mid \sum_i y_i = t] &= \frac{\exp[t\alpha + \sum_{j=1}^p (\sum_{i=1}^N y_i x_{ij})\beta_j] / \prod_i [1 + \exp(\alpha + \sum_{j=1}^p \beta_j x_{ij})]}{\sum_{S(t)} \exp[t\alpha + \sum_{j=1}^p (\sum_{i=1}^N y_i^* x_{ij})\beta_j] / \prod_i [1 + \exp(\alpha + \sum_{j=1}^p \beta_j x_{ij})]} \\ &= \frac{\exp[\sum_{j=1}^p (\sum_{i=1}^N y_i x_{ij})\beta_j]}{\sum_{S(t)} \exp[\sum_{j=1}^p (\sum_{i=1}^N y_i^* x_{ij})\beta_j]} \end{aligned}$$

which no longer depends upon α.

Chapter 14

Poisson Regression

Let $Y \sim Poisson(\mu)$, in which case the pdf is

$$p(y;\mu) = e^{-\mu}\mu^y/y!$$

or

$$p(y;\mu) = exp[y\log(\mu) - \mu - \log(y!)]$$

where $\theta = \log(\mu)$, $a(\phi) = 1$, $b(\theta) = e^\theta$, and $c(y;\phi) = -\log(y!)$.

Consider the case where $Y_1, Y_2, \ldots, Y_N$ are independent Poisson random variables, $Poisson(\mu_i)$ when Y_i denotes the number of events observed in n_i trials and

$$E[Y_i] = \mu_i = n_i\theta_i.$$

Assume that θ_i is dependent upon the explanatory variables x_i and can be modeled as

$$\theta_i = e^{x_i'\beta}$$

in which case,

$$E[Y_i] = \mu_i = n_i e^{x_i'\beta}.$$

Using the natural or canonical link function, the log link, gives

$$\ln\mu_i = \ln n_i + x_i'\beta,$$

where $\ln n_i$ is called the *offset*.

Recall that the relative risk using binary data is the ratio of conditional probabilities. One can define a similar idea with the Poisson model. That is, let $x_{ij} = 0$ if an event is *absent* and $x_{ij} = 1$ if an event is *present*, then the **rate ratio, RR,** is

$$RR = \frac{E[Y_i \mid present]}{E[Y_i \mid absent]} = e^{\beta_j}.$$

The value RR denotes the multiplicative effect of a unit increase in x_j on the rate μ.

14.0.1 Overdispersion for Count Data

Overdispersion occurs when the observed variability in the data is greater than what one would expect based upon the specified model. The Quasi-likelihood approach can be used when encountering over dispersion by letting

$$\nu(\mu) = \phi\mu$$

with the Poisson model or letting

$$\nu(\pi_i) = \phi\pi_i(1-\pi_i)/n_i$$

with the Binomial model where ϕ is a constant > 1. Since, the quasi-likelihood approach is based upon specification of the first two moments one could address over dispersion by specifying a variance which reflects the increased variance. For example, the usual binomial model could be defined using

$$\nu(\pi_i) = \phi\pi_i^2(1-\pi_i^2)/n_i.$$

The value ϕ is cancelled out in the estimating equation, hence, the resultant estimates for β are the same as the original over dispersed model. The difference is found in the standard error of the estimates as, $cov(\beta) = (X'\hat{W}X)^{-1}$ differs by ϕ when using the quasi likelihood adjustment for over dispersion.

An alternative approach when encountering overdispersion in the Poisson loglinear model is to use the Negative Binomial model.

14.1 Negative Binomial GLM

This model is

$$f(y;k,\mu) = \binom{y+k-1}{y}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y, \quad y = 0,1,2,\ldots,$$

where $E(Y) = \mu$ and $var(Y) = \mu + \mu^2/k$. The index k^{-1} is called a *dispersion parameter*. The negative binomial distribution converges to the Poisson, since , $var(Y) \to \mu$, as $k^{-1} \to 0$.

Part V

Other Regression Type Models

Chapter 15

Additional Regression Type Models

In this chapter additional types of regression models are briefly discussed.

15.1 Robust Regression

In this section, regression procedures that are robust to the presence of outliers in the data are introduced. Several different types of procedures are available, however, we will restrict the discussion to M-estimators. These estimators are "maximum likelihood like" estimators. That is, the maximum likelihood estimator (MLE) of β, given by $\hat{\beta}$, maximizes

$$\prod_{i=1}^{n} f(y_i - x_i'\beta),$$

where x_i' is the i^{th} row of X in the model $Y = X\beta + \epsilon$. Equivalently, $\hat{\beta}$ maximizes

$$\sum_{i=1}^{n} \ln f(y_i - x_i'\beta). \tag{15.1}$$

When the data are normally distributed maximizing equation (15.1) is equivalent to the least squares problem of minimizing

$$\sum_{i=1}^{n} (y_i - x_i'\beta)^2 = \sum_{i=1}^{n} \epsilon_i^2.$$

M-estimators are extensions of the basic idea given the above. Let $\rho(u)$ is a specified function of u and s is an estimate of a scale parameter. A robust estimators minimizes

$$\sum_{I=1}^{n} \rho(\epsilon_i/s) = \sum_{i=1}^{n} \rho(\frac{y_i - x_i'\beta}{s}). \tag{15.2}$$

The resultant score equation is

$$\sum_{i=1}^{n} x_{ij}\psi(\frac{y_i - x_i'\beta}{s}) = 0 \tag{15.3}$$

where $\psi(u) = \partial\rho/\partial u$ and x_{ij} is the j^{th} entry of $x_i' = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$. Equation (15.3) will seldom have a close form solution. In which case, solutions can be found using the *Iterative Reweighed Least Squares*

(IRLS). That is, let

$$w_{i\beta} = \frac{\psi[(y_i - x_i'\beta)/s]}{(y_i - x_i'\beta)/s}$$

in which case we have

$$\sum_{i=1}^{n} x_{ij} w_{i\beta}(y_i - x_i'\beta) = 0$$

or

$$\sum_{i=1}^{n} x_{ij} w_{i\beta} y_i = \sum_{i=1}^{n} x_{ij} w_{i\beta} x_i'\beta. \tag{15.4}$$

Equation (15.4) can be written in matrix form as

$$X'W_\beta X\beta = X'W_\beta Y, \tag{15.5}$$

where W_β is a diagonal matrix with elements $w_{i\beta}$. The solution to equation (15.5) is

$$\hat{\beta} = (X'W_\beta X)^{-1} X'W_\beta Y. \tag{15.6}$$

An iterative reweighed least squares estimate for β becomes

$$\hat{\beta}_{q+1} = (X'W_q X)^{-1} X'W_q Y. \tag{15.7}$$

Robust Weight Functions

The weight function used by SAS are given in Figure 15.1

Problems

A number of robust M estimation procedures are given in PROC ROBUSTREG, create an example to illustrate some of these. PROC ROBUSTREG has several robust procedures besides M estimation, select one and give a 10-15 presentation.

15.2 Quantile Regression

Quantile regression generalizes the concept of a univariate quantile to a conditional quantile given one or more covariates.

Definition: 15.1 (Quantile) *The τ^{th} Quantile for the random variable Y with probability distribution function*

$$F(y) = \Pr(Y \leq y)$$

is

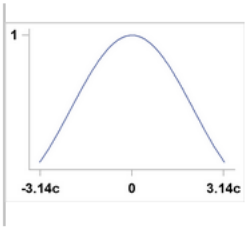
$$Q(\tau) = inf\,\{y : F(y) > \tau\}$$

where $\tau \in [0, 1]$.

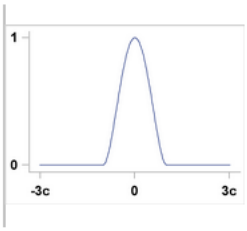
For a random sample $\{y_1, \ldots, y_n\}$ of Y, it is well known that the sample median minimizes the sum of absolute deviations

$$median = \underset{\xi \in R}{\operatorname{argmin}} \sum_{i=1}^{n} \mid y_i - \xi \mid$$

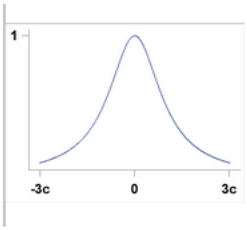
andrews $W(x,c) = \begin{cases} \frac{\sin(\frac{x}{c})}{\frac{x}{c}} & \text{if } |x| \leq \pi c \\ 0 & \text{otherwise} \end{cases}$



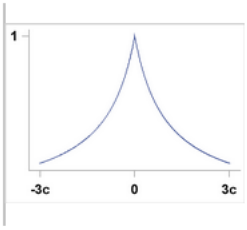
bisquare $W(x,c) = \begin{cases} (1-(\frac{x}{c})^2)^2 & \text{if } |x| < c \\ 0 & \text{otherwise} \end{cases}$



cauchy $W(x,c) = \frac{1}{1+(\frac{|x|}{c})^2}$



fair $W(x,c) = \frac{1}{(1+\frac{|x|}{c})}$



hampel $W(x,a,b,c) = \begin{cases} 1 & |x| < a \\ \frac{a}{|x|} & a < |x| \leq b \\ \frac{a}{|x|}\frac{c-|x|}{c-b} & b < |x| \leq c \\ 0 & \text{otherwise} \end{cases}$

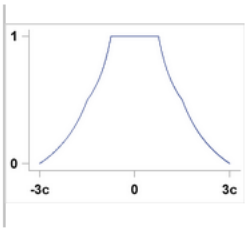


Figure 15.1: SAS Robust Weight Functions

Likewise, the τ^{th} sample quantile $\xi(\tau)$ minimizes

$$\xi(\tau) = \underset{\xi \in R}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi)$$

where $\rho_\tau(z) = z(\tau - I(z < 0))$, $0 < \tau < 1$, and where $I(\cdot)$ denotes the indicator function. The loss function ρ_τ assigns a weight of τ to positive residuals $y_i - \xi$ and a weight of $1 - \tau$ to negative residuals.

Using this loss function, the linear conditional quantile function extends the τ^{th} sample quantile $\xi(\tau)$ to the regression setting in the same way that the linear conditional mean function extends the sample mean. Recall that OLS regression estimates the linear conditional mean function $E(Y \mid X = x) = x'\beta$ by solving for

$$\hat{\beta} = \underset{\beta \in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

The estimated parameter $\hat{\beta}$ minimizes the sum of squared residuals in the same way that the sample mean $\hat{\mu}$ minimizes the sum of squares:

$$\hat{\mu} = \underset{\mu \in R}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mu)^2$$

Likewise, quantile regression estimates the linear conditional quantile function, $Q(\tau|X) = X'\beta(\tau)$, by solving

$$\hat{\beta}(\tau) = \underset{\beta \in R^p}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta)$$

for any quantile $\tau \in (0, 1)$. The quantity $\hat{\beta}(\tau)$ is called the τ^{th} *regression quantile*. The case $\tau = 0.5$, which minimizes the sum of absolute residuals, corresponds to median regression, which is also known as L_1 regression.

The set of regression quantiles $\{\beta(\tau) : \tau \in (0, 1)\}$ is referred to as the *quantile process*.

The PROC QUANTREG computes the quantile function $Q(\tau \mid X)$ and performs statistical inference for the parameters $\beta(\tau)$.

Quantile Regression as an Optimization Problem

The model for linear quantile regression is

$$y = X\beta + \epsilon.$$

L_1 regression, also known as median regression, is a natural extension of the sample median when the response is conditioned on the covariates. In L_1 regression, the least absolute residuals estimate $\hat{\beta}_{LAR}$, referred to as the L_1-norm estimate, is obtained as the solution of the minimization problem

$$\min_{\beta \in R^p} \sum_{i=1}^{n} |y_i - x_i'\beta|$$

More generally, for quantile regression Koenker and Bassett (1978) defined the τ^{th} *regression quantile*, $0 < \tau < 1$, as any solution to the minimization problem

$$\min_{\beta \in R^p} \left[\sum_{i \in \{i: y_i \geq x_i'\beta\}} \tau |y_i - x_i'\beta| + \sum_{i \in \{i: y_i < x_i'\beta\}} (1 - \tau) |y_i - x_i'\beta| \right]$$

The solution is denoted as $\hat{\beta}(\tau)$, and the L_1-norm estimate corresponds to $\hat{\beta}(1/2)$.

$$\min_{\xi \in R} \left[\sum_{i \in \{i: y_i \geq \xi\}} \tau |y_i - \xi| + \sum_{i \in \{i: y_i < \xi\}} (1-\tau)|y_i - \xi| \right].$$

15.3 Classification and Regression Trees

In this section, I have presented some new methods that are being widely used in the era of "Big Data" and "Data Science". Engineers and computer scientists called this area machine learning or "deep learning" for use in problems associated with very large data sets. I have provided several files in the class BOX that provide different aspects of trees and associated methods, such as, random forrest and bagging/boosting methods. As there are many details associated with these methods, I will demonstrate some of the methods as given in a recent JAMA paper[1]. The purpose and the results for the paper are summarized as follows:

- **Context** Estimation of mortality risk in patients hospitalized with **acute decompensated heart failure (ADHF)** may help clinicians guide care.
- **Objective** To develop a practical user-friendly bedside tool for risk stratification for patients hospitalized with ADHF.
- **Design, Setting, and Patients** The Acute Decompensated Heart Failure National Registry (ADHERE) of patients hospitalized with a primary diagnosis of ADHF in 263 hospitals in the United States was queried with analysis of patient data to develop a risk stratification model.[2]
- **Main Outcome Measure** Variables predicting mortality in ADHF.
- **Results** When the derivation and validation cohorts are combined, 37,772 (58%) of 65,275 patient-records had coronary artery disease. Of a combined cohort consisting of 52,164 patient-records, 23,910 (46%) had preserved left ventricular systolic function. In-hospital mortality was similar in the derivation (4.2%) and validation (4.0%) cohorts. Recursive partitioning of the derivation cohort for 39 variables indicated that the best single predictor for mortality was high admission levels of blood urea nitrogen ($\geq$ 43 mg/dL [15.35 mmol/L]) followed by low admission systolic blood pressure ($<$ 115 mm Hg) and then by high levels of serum creatinine ($\geq$ 2.75 mg/dL [243.1 μmol/L]). A simple risk tree identified patient groups with mortality ranging from 2.1% to 21.9%. The odds ratio for mortality between patients identified as high and low risk was 12.9 (95% confidence interval, $10.4 - 15.9$) and similar results were seen when this risk stratification was applied prospectively to the validation cohort.
- **Conclusions** These results suggest that ADHF patients at low, intermediate, and high risk for in-hospital mortality can be easily identified using vital sign and laboratory data obtained on hospital admission. The ADHERE risk tree provides clinicians with a validated, practical bedside tool for mortality risk stratification.

The CART results are given in Figures 15.2.

The decision tree generated by CART analysis of the derivation cohort was tested for its ability to risk stratify patients in the validation cohort. This risk tree was able to stratify patients into high, intermediate, and low risk (Figure 15.2). The mortality OR between the high- and low-risk groups was 10.4 (95% CI,

[1] Risk Stratification for In-Hospital Mortality in Acutely Decompensated Heart Failure – JAMA 2005 vol 293, No. 5 pp 572-580.

[2] The first 33,046 hospitalizations (derivation cohort; October 2001-February 2003) were analyzed to develop the model and then the validity of the model was prospectively tested using data from 32,229 subsequent hospitalizations (validation cohort; March-July 2003). Patients had a mean age of 72.5 years and 52% were female.

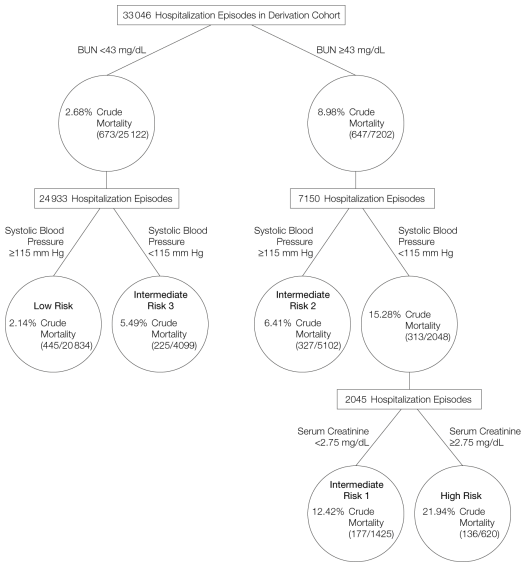


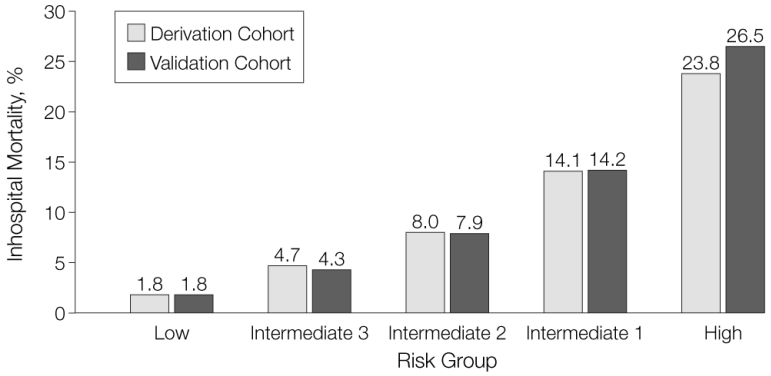
Figure 15.2: Predictors for Mortality using Deviation Cohort

8.4-13.0), with statistically significant differences detected between all risk groups except intermediate risk groups 2 and 3 (Table 15.1). These absolute mortality rates, as well as the clinical characteristics and mortality ORs between risk groups, were similar to those of the derivation cohort (Table 15.1 and 15.2). and comparable risk stratification oc- curred when the analysis was limited to the subset of validation patients with new onset heart failure (in-hospital mortality: 23.6% in the high-risk group; 20.0%, intermediate risk group 1; 5.0%, intermediate risk group 2; 5.1%, intermediate risk group 3; and 1.8%, low- risk group).

Multivariate logistic regression identified BUN level, SBP, heart rate, and age as the most significant mortality risk predictors:

$$\text{log odds of mortality} = 0.0212\ \text{BUN} - 0.0192\ \text{SBP} + 0.0131\ \text{heart rate} + 0.0288\ \text{age} - 4.72. \tag{15.8}$$

The addition of 24 predictors did not meaningfully increase the accuracy of this model. Figure 15.3 compares in hospital mortality rates in the derivation and validation cohorts based on risk groups determined by logistic regression. Based on the area under the receiver operating characteristic curves, the accuracy of the CART model (derivation cohort: 68.7%; validation cohort: 66.8%) was modestly less than that of the more complicated logistic regression model (derivation cohort: 75.9%; validation cohort: 75.7%).



No. of Patients					
Derivation Cohort	20764	4087	5087	1419	617
Validation Cohort	20518	4010	4919	1370	471

Figure 15.3: Predictors for Risk using Deviation and Validation Cohort

15.4 Random Forest

The following material is in a R News article by Andy Liaw and Matthew Wiener[3] A portion of the material has been included here.

Recently there has been a lot of interest in "ensemble learning" – methods that generate many classifiers and aggregate their results. Two well-known methods are boosting (see, e.g., Shapire et al., 1998) and bagging Breiman (1996) of classification trees. In boosting, successive trees give extra weight to points incorrectly predicted by earlier predictors. In the end, a weighted vote is taken for prediction. In bagging, successive trees do not depend on earlier trees – each is independently constructed using a bootstrap sample of the data set. In the end, a simple majority vote is taken for prediction.

Breiman (2001) proposed random forests, which add an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counterintuitive strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against overfitting (Breiman, 2001). In addition, it is very user-friendly in the sense that it has only two parameters (the number of variables in the random subset at each node and the number of trees in the forest), and is usually not very sensitive to their values.

The randomForest package provides an R interface to the Fortran programs by Breiman and Cutler (available at http://www.stat.berkeley.edu/users/breiman/). This article provides a brief introduction to the usage and features of the R functions. Suppose that one has a training data set $d = (X, y)$ where X consists of n observations and p dimensions. y is the dependent variable. If y is continuous then the random forest is regression and if y is categorical, the random forest is for classification. Since the random forest is a CART like procedure with bootstrapping, one needs to specify two parameters, the number of bootstrap samples; $B = ntree$ and the number variables used at each split for each of the bootstrap samples, $m \leq p = mtry$. Note: $m = \sqrt{p}$ or $p/3$ are common values for m.

[3]included in the course BOX.

Table 4. Demographic and Clinical Characteristics of Risk Groups

		Intermediate Risk			
	High Risk*	1†	2‡	3§	Low Risk‖
Derivation Cohort					
Total No. of patients	620	1425	5102	4099	20 834
Age, mean (SD), y	73.6 (12.4)	74.9 (12.1)	73.9 (12.8)	69.7 (15.1)	72.6 (13.9)
No. (%) of patients					
Female	187 (30)	548 (38)	2627 (51)	1761 (43)	11 608 (56)
Coronary artery disease	446 (72)	995 (70)	3365 (66)	2451 (60)	11 770 (57)
Renal insufficiency	496 (80)	720 (51)	3586 (70)	727 (18)	3704 (18)
Diabetes	312 (50)	662 (46)	3043 (60)	1377 (34)	8689 (42)
COPD	156 (25)	452 (32)	1522 (30)	1307 (32)	6420 (31)
No./total (%) of patients with systolic dysfunction¶	400/512 (78)	912/1174 (78)	2007/3906 (51)	2326/3331 (70)	8636/16 761 (52)
Validation Cohort					
Total No. of patients	592	1270	4834	3882	20 820
Age, mean (SD), y	74.0 (12.7)	74.7 (11.6)	74.0 (13.1)	70.0 (15.2)	72.6 (14.1)
No. (%) of patients					
Female	180 (30)	488 (38)	2501 (52)	1660 (43)	11 391 (55)
Coronary artery disease	407 (69)	871 (69)	3002 (62)	2205 (57)	11 236 (54)
Renal insufficiency	479 (81)	650 (51)	3402 (70)	720 (19)	4081 (20)
Diabetes	298 (50)	590 (46)	2887 (60)	1337 (34)	8804 (42)
COPD	160 (27)	406 (32)	1466 (30)	1263 (33)	6436 (31)
No./total (%) of patients with systolic dysfunction	352/472 (75)	774/1054 (73)	1753/3641 (48)	2166/3174 (68)	8115/16 797 (48)

Abbreviation: COPD, chronic obstructive pulmonary disease.
*Defined as blood urea nitrogen level of 43 mg/dL or higher (≥15.35 mmol/L), systolic blood pressure of less than 115 mm Hg, and creatinine level of 2.75 mg/dL or higher (≥243.1 μmol/L).
†Defined as blood urea nitrogen level of 43 mg/dL or higher (≥15.35 mmol/L), systolic blood pressure of less than 115 mm Hg, and creatinine level of less than 2.75 mg/dL (<243.1 μmol/L).
‡Defined as blood urea nitrogen level of 43 mg/dL or higher (≥15.35 mmol/L) and systolic blood pressure of 115 mm Hg or higher.
§Defined as blood urea nitrogen level of less than 43 mg/dL (<15.35 mmol/L) and systolic blood pressure of less than 115 mm Hg.
‖Defined as blood urea nitrogen level of less than 43 mg/dL (<15.35 mmol/L) and systolic blood pressure of 115 mm Hg or higher.
¶Patients had a left ventricular ejection fraction of less than 40% or moderate to severe impairment.

Table 15.1: Breakdown of Risk Groups

The random forests algorithm

1. Draw *ntree* bootstrap samples from the original data.

2. For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample *mtry* of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when $mtry = p$, the number of predictors.)

3. Predict new data by aggregating the predictions of the *ntree* trees (i.e., majority votes for classification, average for regression).

An estimate of the error rate can be obtained, based on the training data, by the following:

1. At each bootstrap iteration, predict the data not in the bootstrap sample (what Breiman calls "out-of-bag", or OOB, data) using the tree grown with the bootstrap sample.

2. Aggregate the OOB predictions. (On the average, each data point would be out-of-bag around 36% of the times, so aggregate these predictions.) Calculate the error rate, and call it the OOB estimate of error rate.

Table 5. In-Hospital Death Between Risk Groups*

	Derivation Cohort		Validation Cohort	
Risk Group Analysis	OR (95% CI)	*P* Value	OR (95% CI)	*P* Value
High vs				
Low	12.9 (10.4-15.9)	<.001	10.4 (8.4-13.0)	<.001
Intermediate 3	4.8 (3.8-6.1)	<.001	4.1 (3.2-5.2)	<.001
Intermediate 2	4.1 (3.3-5.1)	<.001	4.1 (3.3-5.2)	<.001
Intermediate 1	2.0 (1.5-2.5)	<.001	1.6 (1.2-2.1)	<.001
Intermediate 1 vs				
Low	6.5 (5.4-7.8)	<.001	6.5 (5.4-7.8)	<.001
Intermediate 3	2.4 (2.0-3.0)	<.001	2.5 (2.1-3.1)	<.001
Intermediate 2	2.1 (1.7-2.5)	<.001	2.6 (2.1-3.1)	<.001
Intermediate 2 vs				
Low	3.1 (2.7-3.6)	<.001	2.5 (2.2-2.9)	<.001
Intermediate 3	1.2 (1.0-1.4)	.07	1.0 (0.8-1.2)	.94
Intermediate 3 vs low	2.7 (2.2-3.1)	<.001	2.5 (2.2-3.0)	<.001

Abbreviations: CI, confidence interval; OR, odds ratio.
*See Table 4 footnotes for definitions of risk determined by blood urea nitrogen, systolic blood pressure, and creatinine.

Table 15.2: Odds Ratios

Our experience has been that the OOB estimate of error rate is quite accurate, given that enough trees have been grown (otherwise the OOB estimate can bias upward; see Bylander (2002)).

15.5 Example: South African Heart Disease

The data in this example are described in more detail in Hastie and Tibshirani (1987) (HT).

The Coronary Risk-Factor Study (CORIS) baseline survey, carried out in three rural areas of the Western Cape, South Africa (Rousseauw et al., 1983). The aim of the study was to establish the intensity of ischemic heart disease risk factors in that high-incidence region. The data represent white males between 15 and 64, and the response variable is the presence or absence of myocardial infarction (MI) at the time of the survey (the overall prevalence of MI was 5.1% in this region). There are 160 cases in our data set, and a sample of 302 controls.

I have chosen to model the data using CART, Random Forest, and logistic regression (presented in a later chapter). Each of these methods can be preformed using JMP and R. CART and Random Forest is not available using SAS Studio and the SASUniversity Edition.

15.5.1 CART

The JMP output for classification trees is given in Figure (15.4).

From the results of this tree we see that **age**, **famhist**, and **tobaccco** were used for various binary splits. The confusion matrix reveals that 97 subjects with heart disease were misdiagnosed which is probably too high for most applications.

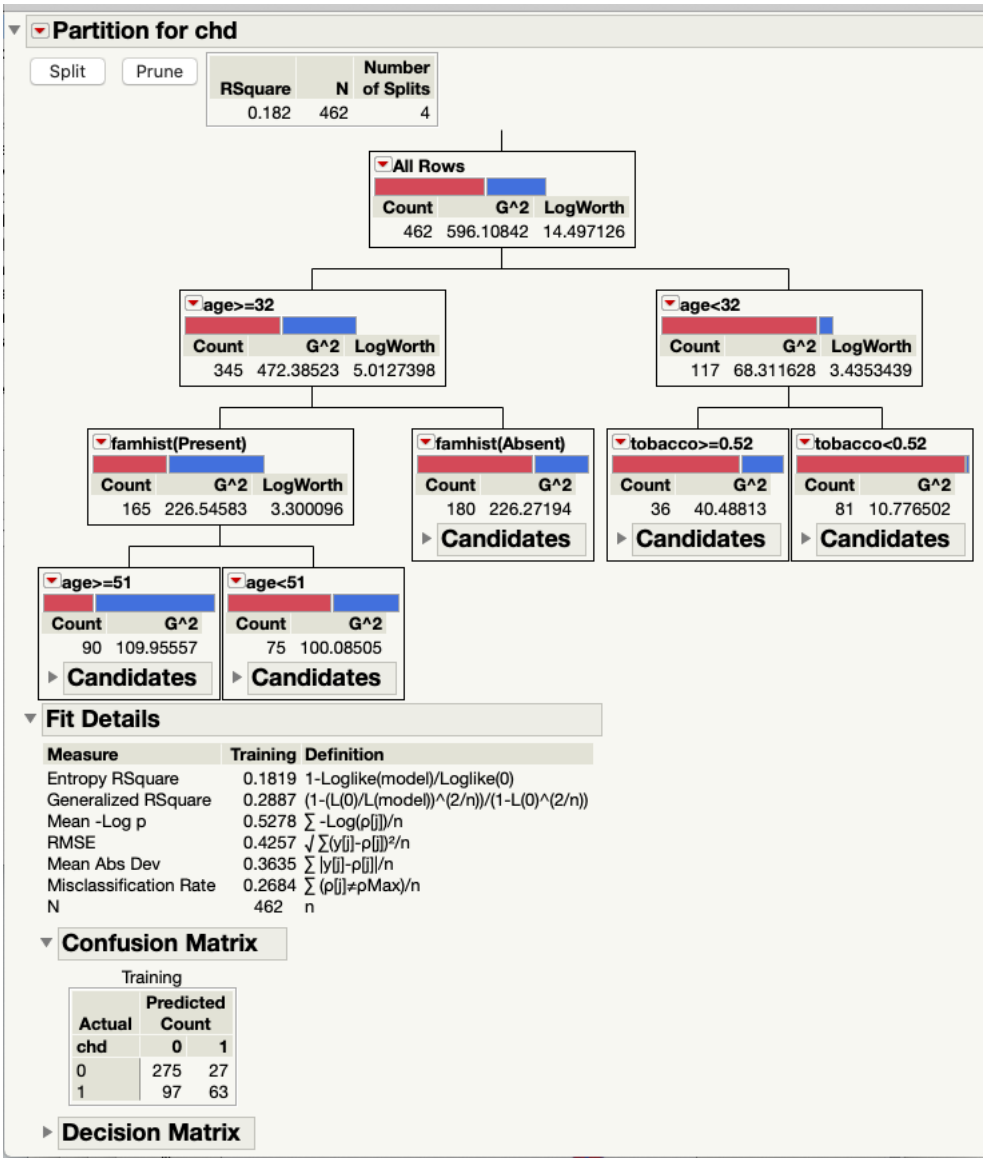


Figure 15.4: Predictors for Heart Disease Using CART

15.5.2 Random Forest

The JMP output for random forest (RF) is given in Figure (15.5).

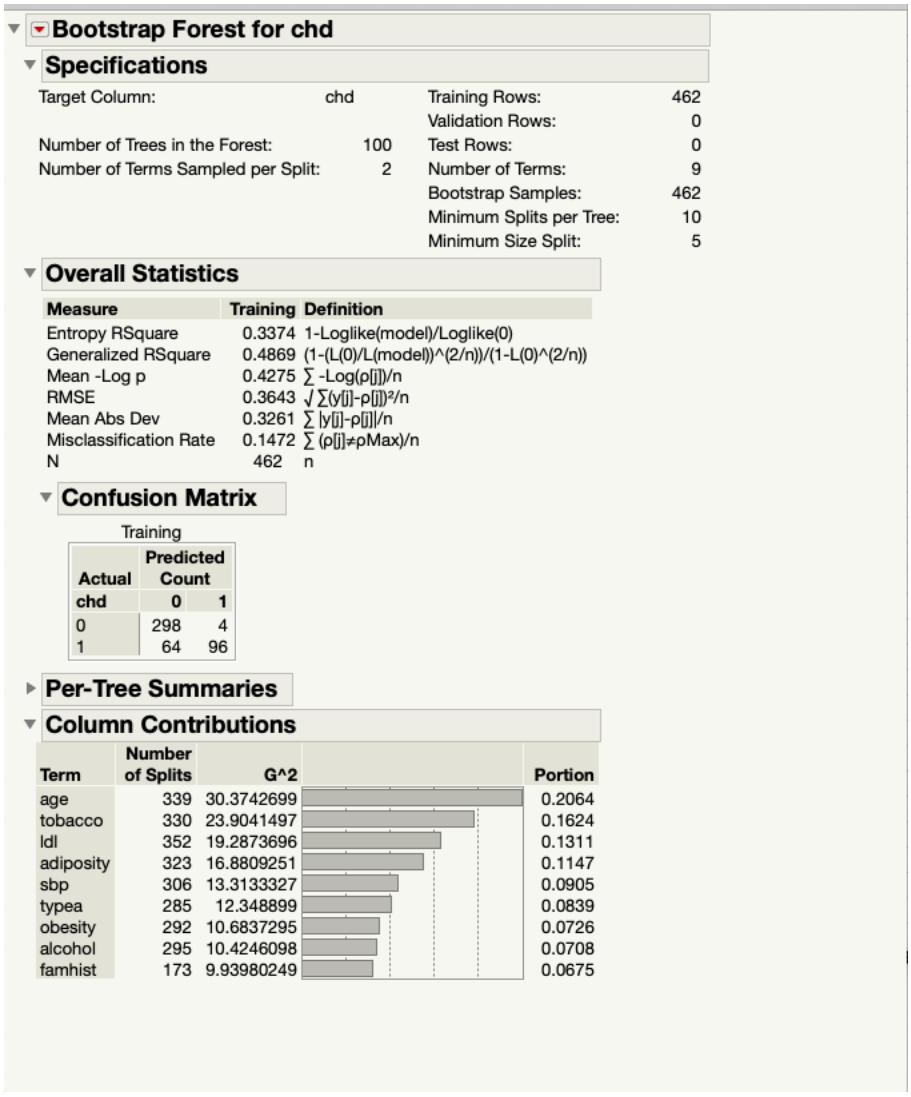
From the results of the RF we see that **age**, **tobaccco**, **ldl**, and **adiposity** were used in more than 10% of the trees. The confusion matrix is better in that 64 subjects with heart disease were misdiagnosed.

15.5.3 Logistic Regression

The JMP output for logistic regression is given in Figure (15.6).

The seen in Figure (15.6) is one of many models that one could try. This one was selected as I could compare these results with those found in Hastie and Tibshirani (1987) page 124 as seen in Figure (15.7). The two tables are very similar (except for the sign of the coefficients where JMP was modeling the probability of not having a heart disease whereas HT was modeling the probability of having a heart disease). We will have a much greater discussion of these models in a later chapter. By using a cutoff at .5, one misclassifies heart disease in 76 subjects using the above variables.

The output from SAS is given in Figure (15.8).



Bootstrap Forest for chd

Specifications

Target Column:	chd	Training Rows:	462
		Validation Rows:	0
Number of Trees in the Forest:	100	Test Rows:	0
Number of Terms Sampled per Split:	2	Number of Terms:	9
		Bootstrap Samples:	462
		Minimum Splits per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Measure	Training	Definition
Entropy RSquare	0.3374	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4869	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4275	∑ -Log(ρ[j])/n
RMSE	0.3643	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.3261	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.1472	∑ (ρ[j]≠ρMax)/n
N	462	n

Confusion Matrix

Training

Actual	Predicted Count	
chd	0	1
0	298	4
1	64	96

Per-Tree Summaries

Column Contributions

Term	Number of Splits	G^2	Portion
age	339	30.3742699	0.2064
tobacco	330	23.9041497	0.1624
ldl	352	19.2873696	0.1311
adiposity	323	16.8809251	0.1147
sbp	306	13.3133327	0.0905
typea	285	12.348899	0.0839
obesity	292	10.6837295	0.0726
alcohol	295	10.4246098	0.0708
famhist	173	9.93980249	0.0675

Figure 15.5: Predictors for Heart Disease Using Random Forest

Generalized Linear Model Fit

Response: chd
Modeling P(chd=0)
Distribution: Binomial
Link: Logit
Estimation Method: Maximum Likelihood
Observations (or Sum Wgts) = 462

Whole Model Test

Model	-LogLikelihood	L-R ChiSquare	DF	Prob>ChiSq
Difference	55.3322795	110.6646	4	<.0001*
Full	242.721931			
Reduced	298.05421			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>ChiSq
Pearson	460.4211	457	0.4463
Deviance	485.4439	457	0.1726

AICc
495.5754

Effect Summary

Source	LogWorth	PValue
age	5.519	0.00000
famhist	4.514	0.00003
tobacco	2.979	0.00105
ldl	2.792	0.00162

Remove Add Edit FDR

Effect Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
tobacco	1	10.736415	0.0011*
ldl	1	9.9415379	0.0016*
famhist	1	17.380821	<.0001*
age	1	21.798674	<.0001*

Parameter Estimates

Term	Estimate	Std Error	L-R ChiSquare	Prob>ChiSq	Lower CL	Upper CL
Intercept	3.7422171	0.4944851	75.536521	<.0001*	2.8092751	4.7521112
tobacco	-0.080701	0.0255148	10.736415	0.0011*	-0.132305	-0.03191
ldl	-0.167584	0.0541898	9.9415379	0.0016*	-0.276193	-0.062947
famhist[Absent]	0.4620583	0.1115915	17.380821	<.0001*	0.2443812	0.6824322
age	-0.044042	0.0097432	21.798674	<.0001*	-0.063544	-0.025263

Studentized Deviance Residual by Predicted

Figure 15.6: Predictors for Heart Disease Using Logistic Regression

TABLE 4.3. *Results from stepwise logistic regression fit to South African heart disease data.*

	Coefficient	Std. Error	Z score
`(Intercept)`	-4.204	0.498	-8.45
`tobacco`	0.081	0.026	3.16
`ldl`	0.168	0.054	3.09
`famhist`	0.924	0.223	4.14
`age`	0.044	0.010	4.52

Figure 15.7: HT Predictors for Heart Disease Using Logistic Regression

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-4.2043	0.4983	71.1732	<.0001
age	1	0.0440	0.00974	20.4333	<.0001
history	1	0.9241	0.2232	17.1448	<.0001
ldl	1	0.1676	0.0542	9.5638	0.0020
tobacco	1	0.0807	0.0255	10.0039	0.0016

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
age	1.045	1.025	1.065
history	2.520	1.627	3.902
ldl	1.182	1.063	1.315
tobacco	1.084	1.031	1.140

Figure 15.8: SAS Predictors for Heart Disease Using Logistic Regression

Part VI

Appendix

Chapter 16

Mathematical Statistics Review

16.1 Background Material

I will briefly review material found in an introductory math stat course so that we might have a common vocabulary with a common notation. If any of this material is new to you, don't panic.

The difficult issue is determining where one begins such a review. Since, statistics is a branch of mathematics one by necessity must know some mathematics. It should be noted that much of what is included in this review is incomplete – not the full story. For it is likened onto the movie entitled "A Few Good Men" when Col. Nathan Jeesep (Jack Nicholson) told JAG Officer Daniel Kaffee (Tom Cruise) that "You can't handle the truth!"[1]. And so it is with you. The truth will be slowly revealed during your time in this program.

16.2 Univariate Distributions

Suppose that one has a population of interest, denoted by Ω, and a member of the population, denoted by ω. Let X denote a function with domain Ω and range $\Re$, the real numbers. That is,

$$X : \omega \in \Omega \to x \in \Re$$

or

$$X(\omega) = x$$

for any $\omega \in \Omega$ and some $x \in \Re$. The function X is said to be a **random variable**[2]. There is a function that is associated with every random variable, X,[3] given by the following definition.

Definition: 16.1 *A* **Cumulative Distribution Function – cdf**, *denoted by $F_X(\cdot)$, for a random variable X exists*[4] *and satisfies*

$$F_X(x) = \Pr[\{\omega \in \Omega : X(\omega) \leq x\}] = \Pr[X \leq x]$$

for every $x \in \Re$.

[1] However, it is not my intent to state incorrect results or ideas for the sake of brevity and/or your inability to handle the truth!

[2] This statement is not completely true as it is neither a variable nor random but rather a measurable function! Does this help? or Do you prefer to not handle the truth!

[3] An additional property of the function X being a random variable.

[4] I have assumed the existence of a probability space, $(\mathcal{X}, \mathcal{A}, \mathcal{P})$.

Definition: 16.2 *A random variable X is said to be* **discrete** *if the range of the function X, denoted by $R_X = Range(X) \subset \Re$, is a countable set.*

Definition: 16.3 *A* **Probability mass function – pmf**, *denoted by $p_X(\cdot)$, for a discrete random variable X exists and satisfies*

$$p_X(x) = \Pr[\{\omega \in \Omega : X(\omega) = x\}] \times I_{R_X}(x) = \Pr[X = x] \times I_{R_X}(x)$$

where R_X is the range for the random variable X and $I_A(a)$ is the indicator function for the set A satisfying $I_A(a) = 1$ if $a \in A$ and $I_A(a) = 0$ if $a \notin A$. Note: $p_X(x)$ is defined for every $x \in \Re$, however, it is only positive when $x \in R_X$, in which case $\Pr[X = x] > 0$.

Definition: 16.4 *A random variable X is said to be* **absolutely continuous (or continuous)** *if there exists a function $f_X(\cdot)$, called a* **probability density function – pdf**, *satisfying*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \qquad \text{for every } x \in \Re.$$

It should be noted that the range of X need not be countable (discrete) nor does the density function $f_X(x)$ need to exist (absolutely continuous). In which case, the random variable is neither discrete nor absolutely continuous and all the properties for X are derived from the cdf, $F_X(x)$.

Definition: 16.5 *The* **Survival function**, *denoted as $S_X(x)$, is*

$$S_X(x) = P(X > x) = 1 - F_X(x) = \int_{x}^{\infty} f_X(t)dt. \tag{16.1}$$

when X is continuous.

It follows that

$$f_X(x) = \frac{dF_X(x)}{dx} = -\frac{dS_X(x)}{dx}. \tag{16.2}$$

Properties of survival curves[5]

Let X and S be defined as above. Then

1. S is real-valued, nonnegative, monotonic, and non-increasing.
2. $S(0) = 1$.
3. $\lim_{x \to \infty} S(x) = 0$.

Family of Distributions Functions

The above notation can be expanded by considering a family of distributions, given by $X \sim F_X(x) \in \Im_\theta$ where $\Im_\theta$ is a family of distributions index by $\theta \in \Theta$. θ is the parameter for the random variable X and indexes the particular member of the family. Θ is the parameter space that contains the set of all possible values of the parameter θ. Note: the parameter, θ, can be a scalar or a vector. Commonly used families of distributions are:

[5] We will not use the survival curve very much in this course, however as the name indicates it is a very important function when modeling survival data.

- *Normal or Gaussian Family:* $X \sim N(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$ and $\theta \in \Theta = \Re \times \Re^+$.
- *Bernoulli:* $X \sim Bernoull(\pi)$ where $\theta = \pi$ and $\Theta = [0, 1] \subset \Re$.
- *Binomial:* $X \sim Binomial(n, \pi)$ where $\theta = (n, \pi) \in \Theta = I^+ \times [0, 1] \subset \Re$.
- *Beta:* $X \sim Beta(\alpha, \beta)$ where $\theta = (\alpha, \beta) \in \Theta = \Re^+ \times \Re^+$.

where $\Re^+$ is the set of positive real numbers and I^+ is the set of positive integers. The product given by $\times$ is the cartesian product of two sets, given by $A \times B = \{(a, b) : a \in A \text{ and } b \in B\}$.
Knowledge of the family and the parameter is sufficient for determining the following properties (adjectives) for the distribution:

Definition: 16.6 *The k^{th}* **unadjusted moment** *for the random variable X is*

$$\mu'_k = E(X^k) = \begin{cases} \sum_{x \in R_X} x^k p_X(x) & \text{when } X \text{ is discrete} \\ \int x^k f_x(x) dx & \text{when } X \text{ is continuous} \end{cases}$$

Definition: 16.7 *The* **population mean** *is the first unadjusted moment, $\mu = \mu'_1 = E(X)$.*

Definition: 16.8 *The k^{th}* **adjusted (centered) moment** *for the random variable X is*

$$\mu_k = E((X - E(X))^k) = \begin{cases} \sum_{x \in R_X} (x - E(X))^k p_X(x) & \text{when } X \text{ is discrete} \\ \int (x - E(X))^k f_x(x) dx & \text{when } X \text{ is continuous} \end{cases}$$

Definition: 16.9 *The* **population variance** *is the second adjusted moment, $\mu_2 = Var(X) = E(X - E(X))^2$.*

Note: the unadjusted and adjusted moments for the random variable X are functions of the parameter $\theta \in \Theta$.

16.2.1 Multivariate Distributions

Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ where X_i is a random variable with parameter θ_i. $\mathbf{X}$ is a p-dimensional random variable with joint CDF given by

$$F_{\mathbf{X}} = F_{X_1, X_2, \ldots, X_p} \in \Im_{\vec{\theta}}, \ \vec{\theta} \in \Theta_p$$

where $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$.

Definition: 16.10 *The random variable* $\mathbf{X}$ *is said to be* **independent** *if and only if*

$$F_{\mathbf{X}} = \prod_{i=1}^{p} F_{X_i}$$

where $F_{X_i} \in \Im_{\theta_i}$.

Definition: 16.11 *The random variable* $\mathbf{X}$ *is said to be* **independent and identically distributed** *if and only if*

$$F_{\mathbf{X}} = \prod_{i=1}^{p} F_{X_i}$$

where $F_{X_i} \in \Im_{\theta}$.

Whenever $p > 1$ other adjectives are needed. Consider the case when $p = 2$ and $\mathbf{X} = (X, Y)$.

Definition: 16.12 *Let X and Y denote two random variables with the joint cdf, $F_{X,Y}(x, y)$ then the $(i, j)^{th}$ joint unadjusted moment is*

$$E(X^i Y^j) = \begin{cases} \sum_{(x,y)\in R_X \times R_Y} x^i y^j p_{X,Y}(x, y) & \text{when } X, Y \text{ are discrete} \\ \int x^i y^j f_{X,Y}(x, y)dxdy & \text{when } X, Y \text{ are continuous} \end{cases}$$

where $p_{X,Y}(x, y) = \Pr[X = x, Y = y]$ is the joint probability mass function.

Definition: 16.13 *Let X and Y denote two random variables with the joint cdf, $F_{X,Y}(x, y)$ then the $(i, j)^{th}$ joint adjusted or centered moment is*

$$E((X-E(X))^i(Y-E(Y))^j) = \begin{cases} \sum_{(x,y)\in R_X \times R_Y} (x - E(X))^i (y - E(Y))^j p_{X,Y}(x, y) & \text{when } X, Y \text{ are discrete} \\ \int (x - E(X))^i (y - E(Y))^j f_{X.Y}(x, y)dxdy & \text{when } X, Y \text{ are continuous} \end{cases}$$

Definition: 16.14 *The* **covariance** *for the random variables, X and Y, is*

$$Cov(X, Y) = E((X - E(X)(Y - E(Y))) = E(XY) - E(X)E(Y).$$

Note: If X, Y are independent then $E(XY) = E(X)E(Y)$ in which case, $Cov(X, Y) = 0$.

16.2.2 Properties of $E(X)$, $Var(X)$ and $Cov(X, Y)$

Let X denote a random variable with expectation $E(X)$ and $Var(X) = E(X - E(X))^2$. For any constants a and b, we have

1. $E(aX \pm b) = aE(X) \pm b$.
2. $Var(aX \pm b) = a^2 Var(X)$.
3. Let $Cov(X, Y)$ denote the covariance of the random variable X and Y, then
   (a) $\text{Cov}(aX \pm b, cY \pm d) = ac\text{Cov}(X, Y)$.
   (b) $-1 \leq \text{Corr}(X, Y) = \rho = \frac{\text{Cov}(X,Y)}{[Var(X)Var(Y)]^{1/2}} \leq 1$.

Linear Combinations

Let $X_1, X_2, \ldots, X_n$ be n random variables and

$$u = \sum_{i=1}^{n} a_i X_i$$

where $E(X_i) = \mu_i$, $Var(X_i) = \sigma_i^2$, and $\text{Cov}(X_i, X_j) = \sigma_{ij}$, then

1. $E(u) = \sum_{i=1}^{n} a_i \mu_i$,
2. $Var(u) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + \sum \sum_{i \neq j} a_i a_j \sigma_{ij}$.

Random Vectors

Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ denote a p-dimensional vector of random variables. Then the expected value of $\mathbf{X}$ is given by $E(\mathbf{X}) = (E(X_1), E(X_2), \ldots, E(X_p))'$. The covariance matrix is an $p \times p$ matrix given by

$$\text{Cov}(\mathbf{X}) = E[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))'] = \Sigma = (\sigma_{ij})$$

where $\sigma_{ij} = \text{Cov}(X_i, X_j)$ and $Var(X_i) = \sigma_{ii} = \sigma_i^2$. Some of the properties for Σ are:

1. Σ is symmetric p.s.d. matrix.
2. $\Sigma = E(\mathbf{X}\mathbf{X}') - E(\mathbf{X})E(\mathbf{X})'$.
3. $\text{Cov}(\mathbf{X} + d) = \text{Cov}(\mathbf{X})$.
4. $tr[\text{Cov}(\mathbf{X})] = trE[(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))'] = E[(\mathbf{X} - E(\mathbf{X}))'(\mathbf{X} - E(\mathbf{X}))] = \sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \sigma_i^2$ is the total variance of $\mathbf{X}$.

   Suppose that A is a $r \times p$ matrix and one defines $\mathbf{V} = A\mathbf{X} \pm B$, then
5. $E(\mathbf{V}) = AE(\mathbf{X}) \pm B$.
6. $\text{Cov}(\mathbf{V}) = A\Sigma A'$. Note $\text{Cov}(\mathbf{V})$ is an $r \times r$ symmetric and at least p.s.d. matrix.

   Suppose that C is a $s \times p$ matrix and one defines $\mathbf{W} = C\mathbf{X} \pm E$, then
7. $\text{Cov}(\mathbf{V}, \mathbf{W}) = A\Sigma C'$. Note $\text{Cov}(\mathbf{V}, \mathbf{W})$ is a $r \times s$ matrix.

16.3 Statistics

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ denote a realization of the random variable $\mathbf{X}$. That is, $\mathbf{X}(\omega_i) = x_i$ for some $\omega_i \in \Omega$.

Definition: 16.15 *The sample observation given by* $\mathbf{x}$ *is said to a* **random sample** *if the joint cdf of* $\mathbf{X}$ *is independent and identically distributed.*

Suppose that the random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is such that $x_i \sim F_x(x) \in \Im_\theta$ with corresponding pdf or pmf given by $f_x(x)$.

Definition: 16.16 *Let* $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ *denote a random sample from a population characterized by* $F_X(x)$*. Then* $T_n = t(X_1, X_2, \ldots, X_n)$ *a real-valued or vectored-valued function of* $\mathbf{x}$ *is said to be a* **statistic** *provided* $(X_1, X_2, \ldots, X_n)$ *is contained in the sample space generated by* $\mathbf{X}$*. Note:* T_n *is a random variable defined on the population* Ω *whose value does not depend upon an unknown parameter* $\theta \in \Theta$*. The c.d.f. for* T_n*, denoted as* $F_{T_n}(t)$*, is called the* **sampling distribution** *for the statistics* T_n*.*

Note: Although the value of T_n is not a function of an unknown θ, the sampling distribution of T_n will be a function of θ and n for any statistic used for the purposes of making inference about the parameter space Θ.

16.3.1 Sampling Distributions

Since the statistic $T_n = t(X_1, X_2, \ldots, X_n)$ is a random variable with c.d.f. $F_{T_n}(t)$ and pdf or pmf $f_{T_n}(t)$, one can compute the moments for T_n as with any random variable. Let $\mu_T = E_\theta(T_n) = g_n(\theta)$ for some function $g_n(\cdot)$. [Again: note that the moments of T_n will depend upon n and θ].

Definition: 16.17 *A statistic* T_n *is said to be* **unbiased (estimator)** *for* $\tau(\theta)$ *for any* $\theta \in \Theta$ *if* $E_\theta(T_n) = g_n(\theta) = \tau(\theta)$*.*

Definition: 16.18 *The variance of T_n is given by $Var(T_n) = E_\theta(T_n - \mu_T)^2$.*

Definition: 16.19 *The standard deviation of T_n, called the* **standard error** *is given by $\sqrt{Var(T_n)}$.*

Definition: 16.20 *If $E_\theta(T_n) = g_n(\theta) \neq \theta$ then T_n is said to be a* **biased estimator for** *θ with the bias given by*

$$Bias_\theta(T_n) = g_n(\theta) - \theta.$$

Definition: 16.21 *The* **Mean Square Error** *of T_n is*

$$MSE_\theta(T_n) = E_\theta(T_n - \theta)^2$$

Deriving Sampling Distributions

You will cover this topic in greater detail in STA 5351 and STA 5352. I have provided a very brief overview using the case where the statistics are continuous. Let $T_n = t(X_1, X_2, \ldots, X_n)$ denote a statistic with cdf $F_{T_n}(t)$ and pdf $f_{T_n}(t)$. The following approaches can be used when determining the distribution and density functions for T_n.

- **Derive the functions using calculus.** For example, suppose that X_i are i.i.d. Uniform(0, θ), $\theta \in \Theta = \Re^+$. In which case, $F_X(x) = \frac{x}{\theta} I_{[0,\theta]}(x) + I_{(\theta,\infty)}(x)$ and $f_X(x) = \frac{1}{\theta} I_{[0,\theta]}(x)$. Define $T_n(t)$ as the n^{th} order statistic given by, $T_n(t) = max(X_1, X_2, \ldots, X_n)$. It follows that,

$$\begin{aligned}
F_{T_n}(t) &= \Pr[T_n \leq t] \\
&= \Pr[max(X_1, X_2, \ldots, X_n) \leq t] \\
&= \Pr[X_1 \leq t; X_2 \leq t; \ldots; X_n \leq t] \\
&= \prod_{i=1}^{n} \Pr[X_i \leq t] \\
&= \prod_{i=1}^{n} \frac{t}{\theta} I_{[0,\theta]}(t) \\
&= [\frac{t}{\theta}]^n \prod_{i=1}^{n} I_{[0,\theta]}(t) \\
&= [\frac{t}{\theta}]^n I_{[0,\theta]}(t) + I_{(\theta,\infty)}(t)
\end{aligned}$$

and

$$\begin{aligned}
f_{T_n}(t) &= \frac{dF_{T_n}(t)}{dt} \\
&= n[\frac{t^{n-1}}{\theta^n}] I_{[0,\theta]}(t).
\end{aligned}$$

It follows that

$$\begin{aligned} E(T_n) &= \int t f_{T_n}(t)dt \\ &= \int_0^\theta n[\frac{t^n}{\theta^n}]dt \\ &= \frac{nt^{n+1}}{(n+1)\theta^n}\Big|_0^\theta \\ &= \left(\frac{n}{n+1}\right)\theta \\ &< \theta, \end{aligned}$$

$$\begin{aligned} Var(T_n) &= E(T_n^2) - E(T_n)^2 \\ &= \int t^2 f_{T_n}(t)dt - \left[\left(\frac{n}{n+1}\right)\theta\right]^2 \\ &= \int_0^\theta n[\frac{t^{n+1}}{\theta^n}]dt - \left[\left(\frac{n}{n+1}\right)\theta\right]^2 \\ &= \frac{nt^{n+2}}{(n+2)\theta^n}\Big|_0^\theta - \left[\left(\frac{n}{n+1}\right)\theta\right]^2 \\ &= \left(\frac{n}{n+2}\right)\theta^2 - \left[\left(\frac{n}{n+1}\right)\theta\right]^2 \\ &= n\left[\frac{[(n+1)^2 - n(n+2)}{(n+1)^2(n+2)}\right]\theta^2 \\ &= \frac{n}{(n+1)^2(n+2)}\theta^2, \end{aligned}$$

and

$$s.e(T_n) = \sqrt{Var(T_n)} = \sqrt{\frac{n}{(n+2)}} \quad \frac{\theta}{(n+1)}.$$

Note: the $Var(T_n) \to 0$ as $n^2 \to \infty$.

Since $E(T_n) < \theta$, T_n is said to be biased. Find a constant (for fixed n) c such that $E(T_n^* = cT_n) = \theta$ and then compute the variance of T_n^*. Note: T_n^* is unbiased for θ.

- **Derive the asymptotic distribution using the Central Limit Theorem.** Whenever the statistic T_n can be written as $\sum_{i=1}^n a_i X_i$ and X_i is such that $E(X_i) = \mu < \infty$ and $Var(X_i) = \sigma^2 < \infty$ then the distribution of T_n can be approximated by the Normal distribution whenever n is large. That is,

$$F_Z(z) \to \Phi(z), \text{ as } n \to \infty$$

where $Z = \frac{T_n - E(T_n)}{\sqrt{Var(T_n)}}$, $E(T_n) = \left(\sum_{i=1}^n a_i\right)\mu$, $Var(T_n) = \left(\sum_{i=1}^n a_i^2\right)\sigma^2$, and $\Phi(\cdot)$ is the c.d.f. for the standard normal. In the above example, another statistic used to estimate θ is $U_n = 2\bar{x} = \frac{2}{n}\sum_{i=1}^n X_i$.

It follows that $Var(U_n) = \frac{\theta^2}{3n} \to 0$ as $n \to \infty$ since $E(X) = \theta/2$ and $Var(X) = \theta^2/12$.

Which of these two estimators, U_n or T_n, would you use, the unbiased one or the biased one? Justify your answer. T_n^* is an unbiased estimator for θ. Which of the two unbiased estimators, U_n or T_n^* would you use? Justify your answer.

- **Derive the standard error using numerical methods, such as, Bootstrapping**. This topic will be covered in much greater detail in an advanced computational statistics course.

  Suppose that one has a realization of a simple random sample of size n, given by $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. A single bootstrap sample given by, $\mathbf{x}_b^* = (x_1^*, x_2^*, \ldots, x_n^*)$, is a sample of size n taken from the above realization when the sampling is done **with replacement**. Suppose that T_n is any statistic. The standard error of T_n can be computed as:

  1. Select B (large number) independent bootstrap samples $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_B^*$, where $\mathbf{x}_b^* = (x_1^*, x_2^*, \ldots, x_n^*)$.
  2. Compute the statistic $T_n = T_n^{*b}$ for each of the $b = 1, 2, \ldots, B$ bootstrap samples.
  3. Estimate the bootstrap standard error by

$$s.e._{boot}(T_n) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(T_N^{*b} - \bar{T_n^*} \right)^2 \right\}^{1/2}$$

  where $\bar{T_n^*} = \frac{1}{B} \sum_{b=1}^{B} T_N^{*b}$.

Bootstrap Example

R provides an easy method of simulating the bootstrap methods. [You will need to install the R package *simple.boot*]. The R code is

```
set.seed(20)
theta = 12  # parameter for the uniform (0, theta)
x <- runif(25)
x = x*theta
b.mean <- one.boot(x, mean, 100)
b.mean$t = 2*b.mean$t
sd(b.mean$t)
boxplot(b.mean$t)

## The statistics is the quantile (1.0 is the max)
b.max <-one.boot(x, quantile, R = 100, probs = 1.0)
sd(b.max$t)
boxplot(b.max$t)
```

The output is

```
> set.seed(20)
> theta = 12
> x <- runif(25)
> x = x*theta
> b.mean <- one.boot(x, mean, 100)
> b.mean$t = 2*b.mean$t
> sd(b.mean$t)
```

```
[1] 1.358726
> boxplot(b.mean$t)
>
> ## The statistics is the quantile (1.0 is the max)
> b.max <-one.boot(x, quantile, R = 100, probs = 1.0)
> sd(b.max$t)
[1] 0.5818098
> boxplot(b.max$t)
```

In this example ($\theta = 12$) the standard error for $2\bar{x}$ is 1.358 (1.92) and the standard error for the n^{th} order statistic is 0.58 (0.43) when $n = 25$ and B = 100. The actual value is in parenthesize.

16.3.2 Sampling Distributions Using the Normal Distribution

Multivariate Normal[6]

The univariate normal density function for X is given by

$$f_X(x) = k\ exp[-1/2\sigma^2(x-\mu)^2]$$

where $E(X) = \mu$, $var(X) = \sigma^2$ and

$$k = (2\pi\sigma^2)^{-1/2}$$

is the normalizing constant. Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ denote an p-dimensional vector with density function given by

$$f(X_1, X_2, \ldots, X_p) = k\ exp[-1/2(\mathbf{X} - E(\mathbf{X}))'\Sigma^{-1}(\mathbf{X} - E(\mathbf{X}))]$$

where,

- $k = (2\pi)^{-p/2} \mid \Sigma \mid^{-1/2}$ is the normalizing constant and $\mid \Sigma \mid$ is the determinate of Σ.
- $E(\mathbf{X}) = \mu = (\mu_1, \mu_2, \ldots, \mu_p)'$ and $\text{Cov}(\mathbf{X}) = \Sigma$.

$\mathbf{X}$ is said to have an p-dimensional multivariate normal distribution with mean, μ and covariance matrix, Σ provided Σ is nonsingular. This is denoted by $\mathbf{X} \sim N_p(\mu, \Sigma)$.

It can be shown that

- $Q = (\mathbf{X} - E(\mathbf{X}))'\Sigma^{-1}(\mathbf{X} - E(\mathbf{X})) \sim \chi^2_p$, where χ^2_p is a Chi-square with p degrees of freedom.
- Suppose that $\mathbf{X} \sim N_p(\mu, \Sigma)$ and A is a $r \times p$. Let $\mathbf{U} = A\mathbf{X} \pm B$ then $\mathbf{U} \sim N_r(\mu_u = A\mu \pm B, \Sigma_u = A\Sigma A')$. The density function for $\mathbf{U}$ exists if $A\Sigma A'$ is nonsingular (i.e. $rank(A) = r$).

16.3.3 Chi-Square, T and F Distributions

This material is presented without derivation in a course such as STA 2381 and is derived in an undergraduate math stat course (e.g., STA 4385). Assume that $z_i \sim N(0, 1)$ are independent for $i = 1, 2, \ldots, n$ then

- $\bar{z} = \frac{1}{n}\sum_{i=1}^n z_i \sim N(0, \sigma^2_{\bar{z}} = \frac{1}{n})$.
- $z_i^2 \sim \chi^2(1)$ and $\sum_{i=1}^n z_i^2 \sim \chi^2(n)$.

[6] This distribution has numerous properties that make is very useful when modeling multivariate data. We will not be doing much with it in this sequence of courses.

- $(n-1)s_z^2 = \sum_{i=1}^n (z_i - \bar{z})^2 \sim \chi^2(n-1)$.
- $\bar{z}$ and s_z^2 are independent.
- If $z \sim N(0,1)$, $u \sim \chi^2(n)$ and z and u are independent then $t = \frac{z}{\sqrt{u/n}} \sim$ t-dist(n).
- If $u \sim \chi^2(n)$, $v \sim \chi^2(m)$, and u and v are independent then $f = \frac{u/n}{v/m} \sim$ F-dist(n,m).
- $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$ then $\mathbf{z}'\mathbf{z} = \sum_{i=1}^n z_i^2 \sim \chi^2(n)$.
- If $x \sim N(\mu, 1)$ then $x^2 \sim \chi^2(df = 1, \lambda = \frac{\mu^2}{2})$. x^2 is said to have a non-central Chi-square distribution with non-centrality parameter λ.

16.3.4 Quadratic Forms of Normal Variables

This material is new and will be used repeatedly in future courses involving linear models. We will use the material when we study ANOVA and linear regression.

1. Let $\mathbf{z} = (z_1, z_2, \ldots, z_n)' \sim N_n(0, I_n)$. Define the quadratic form $q = \mathbf{z}'A\mathbf{z}$. Note: $q = \mathbf{z}'A\mathbf{z} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} z_i z_j$ where $A = (a_{ij})$. It can be shown that
   (a) The expected value of q is $E(q) = tr[A]$.
   (b) The variance of q is $Var(q) = 2\ tr[A^2]$.
   (c) $q \sim \chi^2(a)$ if and only if $A^2 = A$ (A is idempotent) where $a = rank[A] = tr[A]$.
2. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)' \sim N_n(\mu, I_n)$. Define the quadratic form $q = \mathbf{x}'A\mathbf{x}$ then
   (a) The expected value of q is $E(q) = tr[A] + \mu' A\mu$.
   (b) The variance of q is $Var(q) = 2\ tr[A^2] + 4\mu' A^2 \mu$.
   (c) $q \sim \chi^2(a, \lambda)$ if and only if $A^2 = A$ (A is idempotent) where $a = rank[A] = tr[A]$ and $\lambda = 1/2\mu' A\mu$.
   (d) If $\mathbf{x} \sim N_n(\mu, \sigma^2 I_n)$ then $(\mathbf{x}-\mu)'A(\mathbf{x}-\mu)/\sigma^2 \sim \chi^2(a)$ if and only if A is idempotent and $a = tr[A]$.
3. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)' \sim N_n(\mu, V)$ (This means that the $x_i's$ are not independent of one another). Define the quadratic form $q_1 = \mathbf{x}'A\mathbf{x}$ then
   (a) The expected value of q_1 is $E(q_1) = tr[AV] + \mu' A\mu$.
   (b) The variance of q_1 is $Var(q_1) = 2\ tr[AVAV] + 4\mu' AV\mu$.
   (c) $q_1 \sim \chi^2(a, \lambda)$ if and only if $(AV)^2 = AV$ (AV is idempotent) where $a = rank[A]$ and $\lambda = 1/2\mu' A\mu$. Suppose that $q_2 = \mathbf{x}'B\mathbf{x}$ and $\mathbf{t} = C\mathbf{x}$ where C is an $c \times n$ matrix. Then
   (d) $\text{Cov}(q_1, q_2) = 2\ tr[AVBA] + 4\mu' AVB\mu$.
   (e) $\text{Cov}(\mathbf{x}, q_1) = 2\ VA\mu$.
   (f) $\text{Cov}(\mathbf{t}, q_1) = 2\ CVA\mu$.
   (g) q_1 and q_2 are independent if and only if $AVB = BVA = 0$.
   (h) q_1 and $\mathbf{t}$ are independent if and only if $CVA = 0$.
4. (Cochran's Theorem) Let $\mathbf{x} \sim N_n(\mu, V)$. Let A_i be symmetric $n \times n$ matrices with $rank[A_i] = r_i$ for $i = 1, 2, \ldots, m$. Suppose that
$$A = \sum_{i=1}^n A_i$$
with $rank[A] = r$. If AV is idempotent, and $r = \sum r_i$ then $q_i = \mathbf{x}'\mathbf{A_i}\mathbf{x}$ are mutually independent with $q_i \sim \chi^2(df = r_i, \lambda_i = \mu' A_i \mu/2)$.

16.4 Distributions For Categorical Data

Some common categorical or discrete distributions are;

Binomial Distribution

Let $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ denote a realization of n *i.i.d.* Bernoulli trials where $\pi = \Pr[X_i = 1]$ and $1 - \pi = \Pr[X_i = 0]$, then $Y = \sum_{i=1}^{n} X_i$ has a Binomial distribution, denoted by $Bin(n, \pi)$, with probability mass function given by,

$$\Pr[Y = y] = f_Y(y; \pi) = p_Y(y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \quad I_{\{0,1,2,\ldots,n\}}(y)$$

where $I_A(a) = 1$ if $a \in A$ and $I_A(a) = 0$ if $a \notin A$. It can be shown that,

$$E(Y) = n\pi \quad \text{and} \quad Var(Y) = n\pi(1 - \pi).$$

Hypergeometric Distribution

Suppose that one has a finite population of size N that consists of K "successes" and $N - K$ "failures". Let the random variable X_i denote the outcome on the i^{th} trial where the events (success or failure) are selected from the population *without replacement*. Let $Y = \sum_{i=1}^{n} X_i$, where $X_i = 1$ if the outcome is a "success" and $X_i = 0$ if the outcome is a "failure". Y denotes the number of "successes" in the n trials when the events are dependent (sampling without replacement from a finite population). The pmf for Y is given by

$$p_Y(y) = \frac{\binom{K}{y}\binom{N-K}{n-y}}{\binom{N}{n}}$$

for $y = 0, 1, \ldots, n$ provided $0 \leq y \leq K$ and $0 \leq n - y \leq N - K$. It follows that,

$$E(Y) = n(\frac{K}{N}) \quad \text{and} \quad Var(Y) = n(\frac{K}{N})(1 - \frac{K}{N})\frac{N-n}{N-1}.$$

Multinomial Distribution

The Binomial distribution is derived when each random variable (trial) X_i has a binary outcome. Suppose that each trial results in one of c categories. That is, let $x_{ij} = 1$ if trial i has outcome j and $x_{ij} = 0$ otherwise. Then $x_i = (x_{i1}, x_{i2}, \ldots, x_{ic})$ represents a multinomial trial, with $\sum_j x_{ij} = 1$. Let $n_j = \sum_i x_{ij}$ denote the number of outcomes in category j. Then the counts $(n_1, n_2, \ldots, n_c)$ are said to have a multinomial distribution with pdf given by,

$$p(n_1, n_2, \cdots, n_{c-1}) = \left(\frac{n!}{n_1! n_2! \cdots n_c!}\right) \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c} = \frac{n!}{\prod_i n_i!} \prod_i \pi_i^{n_i},$$

where $n = \sum_j n_j$. It follows that,

$$E(n_j) = n\pi_j, \quad Var(n_j) = n\pi_j(1 - \pi_j), \quad \text{and} \quad cov(n_j, n_k) = -n\pi_j\pi_k.$$

Poisson Distribution

The Poisson distribution has pdf,

$$\Pr[Y = y] = f_Y(y; \lambda) = p(y) = e^{-\lambda}\lambda^y / y! \quad I_{\{0,1,2,\ldots,\}}(y).$$

It can be shown that, $E(y) = Var(y) = \lambda$.

Connection between the Poisson and the Multinomial Distribution

In the above Poisson distribution the random variable $Y = y = n$ is random. Suppose that n is fixed and that each n_i has a Poisson distribution then the random variables Y_i are no longer independent. In which case we have,

$$\begin{aligned}\Pr[Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c \mid \sum_i Y_i = n] &= \frac{\Pr[Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c]}{\Pr[\sum_i Y_i = n]} \\ &= \frac{\prod_i [exp(-\lambda_i)\lambda_i^{n_i}/n_i!]}{exp(-\sum_i \lambda_i)(\sum_i \lambda_i)^n/n!} \\ &= \frac{n!}{\prod_i n_i!}\prod_i \pi_i^{n_i}, \end{aligned} \tag{16.3}$$

where $\pi_i = \lambda_i/(\sum_i \lambda_i)$.

The above distributions are counting distributions where one is "interested" in the number of "successes" observed. The following distributions are useful in what is called the *inverse sampling* or *waiting time* problem where one is interested in the number of events (time) needed to observe a specified number of "successes".

Geometric Distribution

Suppose that one has a sequence of Bernoulli trials with $X_i \sim Bernoulli(p)$ where $p = \Pr[\text{"}success\text{"}]$ then the pmf for the random variable Y is given by

$$p_Y(y) = q^{y-1}p$$

where $y = 1, 2, \ldots$ and $q = 1 - p$. Y is the number of trials needed to observed a "success". Another representation is found by letting $Z = Y - 1$ denote the number of "failures" observed prior to the first "success". The pmf for Z is given by

$$p_Z(z) = q^z p$$

where $z = 0, 1, \ldots$. Recalling that the geometric series satisfies,

$$\sum_{j=0} r^j = \frac{1}{1-r}$$

provided $|r| < 1$. It follows that

$$E(Y) = \frac{1}{p} \quad \text{and} \quad Var(Y) = \frac{q}{p^2}$$

and

$$E(Z) = E(Y) - 1 = \frac{1}{p} - 1 = \frac{q}{p}.$$

Negative Binomial Distribution

Let $W = \sum_{i=1}^r Y_i$ where $Y_i \sim Geometric(p)$. That is, W is the waiting time for the r^{th} "success". The pmf for W is given by

$$p_W(w) = \binom{w-1}{r-1} p^r q^{w-1}$$

where $w = r, r+1, \ldots$. Since each Y_i is assumed to be independent it follows that

$$E(W) = rE(Y) = \frac{r}{p} \quad \text{and} \quad Var(W) = rVar(Y) = \frac{rq}{p^2}.$$

Relationship Between the Negative Binomial and the Poisson

If one parameterizes the model where $E(W) = \mu = \frac{r}{p}$ and $Var(W) = \mu + \frac{\mu^2}{k}$ the pmf for W can be written as

$$f_W(w) = \binom{w+k-1}{w}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^w, \quad w = 0, 1, 2, \ldots,$$

where k and μ are the new parameters. This distribution has

$$E(W) = \mu \quad \text{and} \quad var(W) = \mu + \mu^2/k.$$

The index k^{-1} is called a *dispersion parameter*. As $k^{-1} \to 0, var(W) \to \mu$ and the negative binomial distribution converges to the Poisson with parameter $\lambda = \mu$.

16.5 Matrices

The material in this section is intended as a review of material that you should have seen in an undergraduate math course in linear algebra.

Matrix Algebra - Review

A matrix $\mathbf{A} = (a_{ij}), i = 1, 2, \ldots, r, j = 1, 2, \ldots, c$ is said to be an $r \times c$ matrix given by

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1c} \\ a_{21} & a_{22} & \ldots & a_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ a_{r1} & a_{r2} & \ldots & a_{rc} \end{pmatrix}$$

A vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is said to be a $n \times 1$ row vector, x' is a $1 \times n$ column vector given by

$$\mathbf{x}' = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

16.5.1 Special Matrices

1. $\mathbf{D} = diag(A)$ is the diagonal of the $r \times r$ matrix A given by

$$\mathbf{D} = \begin{pmatrix} a_{11} & 0 & \ldots & 0 \\ 0 & a_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & a_{rr} \end{pmatrix}.$$

2. $\mathbf{I}_n$ is called the $n \times n$ identity matrix given by

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

3. $\mathbf{J}_n$ is an $n \times n$ matrix with each element equal to one.

4. $\mathbf{j}$ is a $n \times 1$ vector with each element equal to one where $\mathbf{J} = \mathbf{j}\mathbf{j}'$.

16.5.2 Addition

$C = A \pm B$ is defined as $c_{ij} = a_{ij} \pm b_{ij}$ provided both A and B have the same number of rows and columns. It can easily be shown that $(A \pm B) \pm C = A \pm (B \pm C)$ and $A + B = B + A$.

16.5.3 Multiplication

$C = AB$ is defined as $c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$ provided A and B are conformable matrices (A is $r \times p$ and B is $p \times c$). Note: Even if both AB and BA are defined they are not necessarily equal. It follows that $A(B \pm C) = AB \pm AC$). Two vectors a and b are said to be orthogonal, denoted by $a \perp b = 0$, if $ab = \sum_{i=1}^{n} a_i b_i = 0$.

16.5.4 Kronecker or Direct Product

If A is $m \times n$ and B is $s \times t$, the *direct or Kronecker product* of A and B, denoted by $A \otimes B$, is an $ms \times nt$ matrix given by

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix}.$$

Properties are given as

1. $(A \otimes B)(C \otimes D) = (AC \otimes BD)$.
2. $((A + B) \otimes (C + D)) = (A \otimes C) + (A \otimes D) + (B \otimes C) + (B \otimes D)$.
3. $A \otimes (B \otimes C) = (A \otimes B) \otimes C$.

16.5.5 Inverse

A $n \times n$ matrix A is said to be nonsingular if there exists a matrix B satisfying $AB = BA = I_n$. B is called the inverse of A is denoted by A^{-1}.

16.5.6 Transpose

If A is $r \times c$ then the transpose of A, denoted by A', is a $c \times r$ matrix. It follows that

1. $(A')' = A$
2. $(A \pm B)' = A' \pm B'$
3. $(AB)' = B'A'$
4. If $A = A'$ then A is said to be symmetric.
5. $A'A$ and AA' are symmetric.
6. $(A \otimes B)' = (A' \otimes B')$.

16.5.7 Trace

Definition: 16.22 *Suppose that the matrix* $A = (a_{ij}), i = 1, \ldots, n, j = 1, \ldots, n$ *then the trace of* A *given by* $tr[A] = \sum_{i=1}^{n} a_{ii}$.

Provided the matrices are conformable

1. $tr[A] = tr[A']$.
2. $tr[A \pm B] = tr[A] \pm tr[B]$.
3. $tr[AB] = tr[BA]$.
4. $tr[ABC] = tr[CAB] = tr[BCA]$.
5. $tr[A \otimes B] = tr[A]tr[B]$.

   For a square matrix A, one can write $Ax = \lambda x$ for some non-null vector x, then λ is called a *characteristic or eigenvalue or latent root* of A. x is called the corresponding characteristic vector (eigenvector or latent vector).

   If A is a symmetric $n \times n$ matrix with eigenvalues λ_i for $i = 1, 2, \ldots, n$, then
6. $tr[A] = \sum_{i=1}^{n} \lambda_i$
7. $tr[A^s] = \sum_{i=1}^{n} \lambda_i^s$
8. $tr[A^{-1}] = \sum_{i=1}^{n} \lambda_i^{-1}$, A nonsingular.

16.5.8 Rank

Suppose that A is a $r \times c$ matrix with r rows $a_1, a_2, \ldots, a_c$ are said to be linearly independent if no a_i can be expressed as a linear combination of the remaining $a_i's$, that is, there does not exist a non-null vector $c = (c_1, c_2, \ldots, c_r)$ such that $\sum_{i=1}^{r} c_i a_i = 0$. It can be shown that the number of linearly independent rows is equal to the number of linearly independent columns of any matrix A and that number is the rank of the matrix. If the rank of A is r then the matrix A is said to be full row rank. If the rank of A is c then A is said to be full column rank.

1. $rank[A] = 0$ if and only if $A = 0$.
2. $rank[A] = rank[A']$.
3. $rank[A] = rank[A'A] = rank[AA']$.
4. $rank[AB] \leq min\{rank[A], rank[B]\}$
5. If A is any matrix, and P and Q are any conformable nonsingular matrices then $rank[PAQ] = rank[A]$.
6. If A is $r \times c$ with rank r then AA' is nonsingular ($(AA')^{-1}$ exists and $rank[AA'] = r$). If the rank of A is c then $A'A$ is nonsingular ($(A')^{-1}$ exists and $rank[A'A] = c$).
7. If A is symmetric, then $rank[A]$ is equal to the number of nonzero eigenvalues.

16.5.9 Quadratic Forms

Let $\mathbf{A}$ be a symmetric $n \times n$ matrix and $\mathbf{x} = (x_1, x_2, \ldots, n)$ be a vector. Then $q = \mathbf{x}'\mathbf{A}\mathbf{x}$, is called a quadratic form of A. The quadratic form is a second degree polynomial in the $x_i's$, since $q = \mathbf{z}'A\mathbf{z} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}z_iz_j$ where $A = (a_{ij})$. In this definition I have assumed that $\mathbf{A}$ is symmetric. This follows since any non-symmetric matrix $\mathbf{B}$ can be written as $1/2[\mathbf{B}+\mathbf{B}']$ and $1/2[\mathbf{B}+\mathbf{B}']$ is symmetric. Furthermore, $\mathbf{x}'\mathbf{B}\mathbf{x} = 1/2[\mathbf{x}'\mathbf{B}\mathbf{x} + \mathbf{x}'\mathbf{B}'\mathbf{x}]$.

16.5.10 Positive Semidefinite Matrices

A symmetric matrix A is said to be positive semidefinite (p.s.d.) if and only if $q = x'Ax \geq 0$ for all x.

1. The eigenvalues of p.s.d. matrices are nonnegative.
2. If A is p.s.d. then $tr[A] \geq 0$.
3. A is p.s.d. of rank r if and only if there exists an $n \times n$ matrix R of rank r such that $A = RR'$.
4. If A is an $n \times n$ p.s.d. matrix of rank r, then there exists an $n \times r$ matrix S of rank r such that $S'AS = I_r$.
5. If A is p.s.d., then $X'AX = 0 \Rightarrow AX = 0$.

16.5.11 Positive Definite Matrices

A symmetric matrix A is said to be positive definite (p.d.) if and only if $q = x'Ax > 0$ for all $x, x \neq 0$.

1. The eigenvalues of p.d. matrices are positive.
2. A is p.d. if and only if there exists an nonsingular matrix R such that $A = RR'$.
3. If A is p.d. then so is A^{-1}.
4. If A is p.d. then $rank[CAC'] = rank[C]$.
5. If A is $n \times n$ p.d. matrix and C is a $p \times n$ matrix of rank p, then CAC' is p.d.
6. If X is $n \times p$ of rank p then $X'X$ is p.d.
7. If A is p.d. if and only if all the leading minor determinants of A are positive.
8. The diagonal elements os a p.d. matrix are all positive.
9. (Cholesky decomposition). Is A is p.d. there exists a unique upper triangular matrix U with positive diagonal elements such that $A = U'U$.

16.5.12 Idempotent Matrices

A matrix P is said to be idempotent if $P^2 = P$. A symmetric idempotent matrix is called a projection matrix.

1. If P is symmetric, then P is idempotent and of rank r if and only if it has r eigenvalues equal to unity and $n - r$ eigenvalues equal to zero.
2. If P is a projection matrix then the $tr[P] = rank[P]$.
3. If P is idempotent, so is $I - P$.
4. Projection matrices are positive semidefinite.

16.5.13 Orthogonal Matrices

An $n \times n$ matrix A is said to be orthogonal if and only $A^{-1} = A'$. If A is orthogonal then

1. $-1 \leq a_i \leq 1$.
2. $AA' = A'A = I_n$.
3. $\mid A \mid = 1$.

16.5.14 Vector Differentiation

Let X be an $n \times m$ matrix with elements x_{ij}, then if $f(X)$ is a function of the elements of X, we define

$$\frac{df}{dX} = [(\frac{df}{dx_{ij}})]$$

then

1. $\frac{d(\beta' a)}{d\beta} = a$.
2. $\frac{d(\beta' A\beta)}{d\beta} = 2A\beta$. ($A$ symmetric).
3. if $f(X) = a'Xb$, then $\frac{df}{dX} = ab'$.
4. if $f(X) = tr[AXB]$, then $\frac{df}{dX} = A'B'$.
5. if X is symmetric and $f(X) = a'Xb$, then $\frac{df}{dX} = ab' + b'a - diag(ab')$.
6. if X is symmetric and $f(X) = tr[AXB]$, then $\frac{df}{dX} = A'B' + BA - diag(BA)$.
7. if X and A are symmetric and $f(X) = tr[AXAX]$, then $\frac{df}{dX} = 2AXA$.

16.5.15 The Generalized Inverse

A matrix B is said to be the generalized inverse of A if it satisfies $ABA = A$. The generalized inverse of A is denoted by A^-. If A is nonsingular then $A^{-1} = A^-$. If A is singular then A^- exists but is not unique.

1. If A is an $r \times c$ matrix of rank c. Then the generalized inverse of A is $A^- = (A'A)^{-1}A'$.
2. If A is an $r \times c$ matrix of rank r. Then the generalized inverse of A is $A^- = A(AA')^{-1}$.
3. If A is an $r \times c$ matrix of rank c. Then $A(A'A)^-A'$ is symmetric, idempotent, of rank A, and unique.

16.5.16 Generalized Inverse of $X'X$

Let G denote the generalized inverse of $X'X$, that is

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

Clearly, $X'X$ is symmetric although G may not be. However, it follows that G' is also the generalized inverse of $X'X$, or

$$X'XG'X'X = X'X,$$

and that

$$(X'X)^- = GX'XG'$$

which is symmetric.
Other properties of G are;

- G' is also the generalized inverse of $X'X$.
- $XGX'X = X$ or GX' is the generalized inverse of X.
- $X'XG'X' = X'$ or XG' is the generalized inverse of X'.
- XGX' is invariant to the choice of G.
- XGX' is symmetric for any choice of G.
- For V being symmetric and positive definite (i.e. a covariance matrix) then

$$X(X'V^{-1}X)^{-}X'V^{-1} \text{ is invariant to } (X'V^{-1}X)^{-}$$

and

$$X(X'V^{-1}X)^{-}X'V^{-1}X = X.$$

16.5.17 Solution of Linear Equations

A system of linear equations given by $Ax = b$ is said to be consistent and has a solution which can be expressed as $\tilde{x} = A^{-}b$. If A is nonsingular then $\tilde{x}$ is unique.