Rug 61139

2. CASO 2
EFECTOS DE LA APLICACION DE AGUA EN EXCESO SOBRE EL ARE Y LA PRODUCCION DE CAÑA. EL CASO DE UN INGENIO AZUCARERO

2.1. INTRODUCCION

El siguiente es un caso que presenta cómo el análisis de registros sobre manejo de aguas y fertilizante en uno de los ingenios del sector permite, mediante la utilización de la técnica de los modelos lineales, mostrar la posibilidad de incrementar sustancialmente la eficiencia al disminuir los costos de producción en la práctica de riegos y que trae como consecuencia positiva la preservación del recurso hídrico que desafortunadamente se está agotando.

La teoría de los modelos lineales conjuntamente con la estimación de contrastes de efectos simples de tratamientos, la estimación de promedios por mínimos cuadrados y los métodos de regresión lineal simple, permiten evaluar y caracterizar tendencias de los diferentes factores bajo estudio.

El propósito de este trabajo es cuantificar el efecto de la aplicación de nitrógeno y de agua en exceso sobre el azúcar recuperada (ARE) y la producción de caña de azúcar de las suertes cultivadas a nivel comercial, durante el ciclo de cultivo y valorar en términos económicos la práctica de riego en condiciones de exceso.

Es importante aclarar que la información analizada no atiende a una planeación, ni a un diseño de experimento como tal; por lo contrario, son datos que corresponden a las cantidades reales de insumos aplicados en 873 suertes cosechadas comercialmente en las tierras propias y en participación de un ingenio azucarero del Valle del Cauca durante 1991-1993. Al encontrar tendencias o comportamientos significativos de las variables de producción ante los cambios en la aplicación de

nitrógeno y en el agua en exceso, hace pensar que las conclusiones a que se llegan en este análisis son válidas y con seguridad en un estudio planeado sobre respuestas de producción controlando la aplicación de los insumos involucrados conjuntamente, se pueden llegar a encontrar resultados mucho más alarmantes sobre la incidencia de la aplicación de insumos excesivamente.

2.2. METODOLOGIA

Con base en las mediciones comerciales del total de agua aplicada en los riegos y la precipitación total se establecen los niveles de exceso de agua que las suertes efectivamente presentaron en el transcurso del ciclo del cultivo[1], definidos como la diferencia entre el total de agua y lo requerido para el desarrollo del cultivo de caña de azúcar en todo el ciclo. Así, el 98,5% de las suertes con las tres principales variedades: CP 57-603, MZC 74-275 y PR 61-632 que equivalen a 691 suertes (11.520 hectáreas) cosechadas en los últimos tres años por este ingenio presentaron aplicación de agua en exceso; excesos que se distribuyen en 8 categorías de 3000 m^3 de agua por ciclo, que van desde cero m^3 hasta niveles que sobrepasan los 18.000 m^3 y es por tal motivo que se considera de fundamental importancia analizar cómo este factor incide sobre las variables de producción y cuantificar su efecto de manera aislada y en forma conjunta con otros factores y con el nitrógeno aplicado.

Se caracterizan las respuestas del tonelaje de caña por hectárea (TCH) y del azúcar recuperable (ARE) debidas a variedades, grupos de exceso de agua, grupos de corte y sus interacciones y a la variable dosis de nitrógeno por hectárea considerada como una covariable, mediante el siguiente modelo de 3 factores.

$$Y_{ijkl} = M + V_i + E_j + VE_{ij} + C_k + VC_{ik} + EC_{jk} + VEC_{ijk} + \beta X_{ijkl} + E_{ijkl}$$

[1] "La precipitación efectiva se define como el 80% del total de precipitación ocurrida en el trascurso del ciclo de cultivo; el cultivo requiere en total 12000 M^3*ha^{-1} de agua por ciclo, donde el agua de riego y la precipitación deben establecer el balance hídrico que permita que la planta obtenga todo el producto potencial; así que, se considera agua en exceso a los valores del total de agua recibida por una hectárea por ciclo que están por encima de 12000 M^3": Programa de Agronomía de CENICAÑA.

Donde

Y_{ijkl}: Variable de respuesta correspondiente a la ijkl ésima unidad de estudio

M: Efecto de media general

V_i: Efecto de la variedad i (i=1,2,3)

E_j: Efecto del grupo de exceso j (j=1,2,3,4,5)

C_k: Efecto del grupo de corte k (k=1,2)

VE_{ij}: Efecto de la interacción de variedad-grupo de exceso

VC_{ik}: Efecto de la interacción de variedad-grupo de corte

EC_{jk}: Efecto de la interacción de grupo exceso-grupo corte

VEC_{ijk}: Efecto de la interacción variedad-grupo exceso-grupo corte

X_{ijkl}: Cantidad fija observada de dosis de nitrógeno/ha

ß: Coeficiente de regresión

E_{ijkl}: "Error aleatorio"

De los 8 grupos de exceso de agua antes mencionados se consideraron solamente 5 grupos debido a que ellos concentraron el 93.1% de la información.

El **Cuadro 1** muestra las variables independientes y los niveles de los factores considerados. Se utilizó el programa SAS para el análisis de la información.

CUADRO 1. Descripción de variables.

Variables Dependientes:	TCH y RDTO
Variables Independientes (Factores):	
Variedad:	
	CP 57-603
	MZC 74-275
	PR 61-632
Grupos de Exceso de Agua:	
	1. (0-3000]
	2. (3000-6000]
	3. (6000-9000]
	4. (9000-12000]
	5. (12000-15000]
Grupos de Corte:	
	1. Plantillas
	2. Socas
Nitrógeno por Hectárea:	
	79-186Kg * Ha^{-1}

2.3. RESULTADOS

El análisis de varianza para el modelo, mediante la utilización del procedimiento GLM del programa estadístico SAS, y tomando de la suma de cuadrados tipo III, dado que se está en presencia de un cuasiexperimento desbalanceado con 3 factores en el que existe por lo menos un dato en cualquiera de las celdas que conforman la combinación de niveles de los factores, permitió detectar que hay un efecto significativo de la aplicación de nitrógeno por hectárea tanto en la variable TCH como en el ARE; de tal manera que con la opción SOLUTION del procedimiento se pudo estimar que por cada incremento de una unidad de nitrógeno por hectárea se encuentra asociado de manera lineal un incremento promedio de 0.09 toneladas de caña por hectárea y una disminución del ARE de 0.0063 unidades. Vale la pena aclarar que el rango de variación del nitrógeno por hectárea estuvo entre 79 y 186 $kg*ha^{-1}$ y que su efecto es independiente de los otros factores, tal como es la variedad, los grupos de exceso de agua, etc, al tenerse en cuenta que no hubo suficiente evidencia para afirmar lo contrario (No hubo interacciones significativas del nitrógeno por hectárea con los otros factores).

El análisis de varianza también detectó como significativa la interacción variedad-grupo de excesos, y no hubo suficiente evidencia para decir que las interacciones de variedad - grupo de corte y variedad - grupo de exceso - grupo de corte son significativas, esto quiere decir que las variedades responden de diferente manera según la cantidad de agua que se le aplique. En el **Cuadro 2** se muestran los promedios de TCH y ARE asociados a cada grupo de exceso de agua para cada una de las variedades.

CUADRO 2. Respuestas de la producción de caña y del azúcar recuperable de las variedades según grupos de exceso de agua recibida.

Variedad	Grupo de Exceso de Agua (m^3)	Promedio de TCH	Promedio de ARE
CP 57-603	**1.** (0 -3000]	116,3 b[1)]	11,31 a
	2. (3000-6000]	119,1 b	11,52 a
	3. (6000-9000]	117,9 b	11,30 a
	4. (9000-12000]	124,8 a	11,24 a
	5. (12000-15000]	121,3 a	11,47 a
MZC 74-275	**1.** (0-3000]	139,6 b	11,35 a
	2. (3000-6000]	139,4 b	11,42 a
	3. (6000-9000]	149,2 a	11,31 a
	4. (9000-12000]	146,8 a	11,24 a
	5. (12000-15000]	148,3 a	11,41 a
PR 61-632	**1.** (0-3000]	151,7 b	10,90 a
	2. (3000-6000]	161,5 a	10,76 a
	3. (6000-9000]	162,0 a	10,62 a
	4. (9000-12000]	163,6 a	10,86 a
	5. (12000-15000]	167,0 a	10,81 a

1) Promedios con la misma letra no son significativamente diferentes (P=0.10)

Se observa que para la variedad CP 57-603 se obtienen diferencias significativas en peso en caña para los grupos de exceso de agua por debajo de 9000 m^3 al ser comparados con los grupos de exceso que

están por encima de este valor. Para la MZC 74-275, estas diferencias se dan siginificativamente en los niveles por debajo de 6000 m^3 comparados con los resultados obtenidos en niveles de exceso superiores a éstos; mientras que la PR 61-632 el valor de diferencia está determinado por 3000 m^3 de agua en exceso. Las tres variedades responden de diferente manera a la aplicación de agua en exceso en cuanto a la variable TCH, pero sobre la variable ARE no tuvo ningún efecto significativo.

No sólo es importante determinar que existe un efecto significativo de los grupos de exceso de agua sino que es necesario determinar cómo es su comportamiento, con este fin se hizo un análisis de varianza descomponiendo las sumas de cuadrados debidos a los grupos de exceso y su interacción con variedades en componentes lineales, cuadráticos, cúbicos y de grado cuatro, el cual permitió caracterizar una respuesta lineal significativa ante el incremento del agua en exceso para cada variedad.

La estimación de los coeficientes de regresión lineal permite mostrar que por cada incremento de 1000 m^3 de agua en exceso que recibe una hectárea con caña de azúcar durante su período de cultivo con respecto a lo requerido, la variedad CP 57-603 tiene una productividad marginal de 0,5 toneladas, la MZC 74-275 de 0,83 toneladas y la PR 61-632 de 1,1 toneladas de caña por hectárea.

En los **Gráficos 1, 2 y 3** se muestra el comportamiento para cada una de las variedades, la tendencia en producción de caña por hectárea cosecha se expresa como una función del exceso de agua, medido en 1000 m^3 por hectárea durante el ciclo de cultivo.

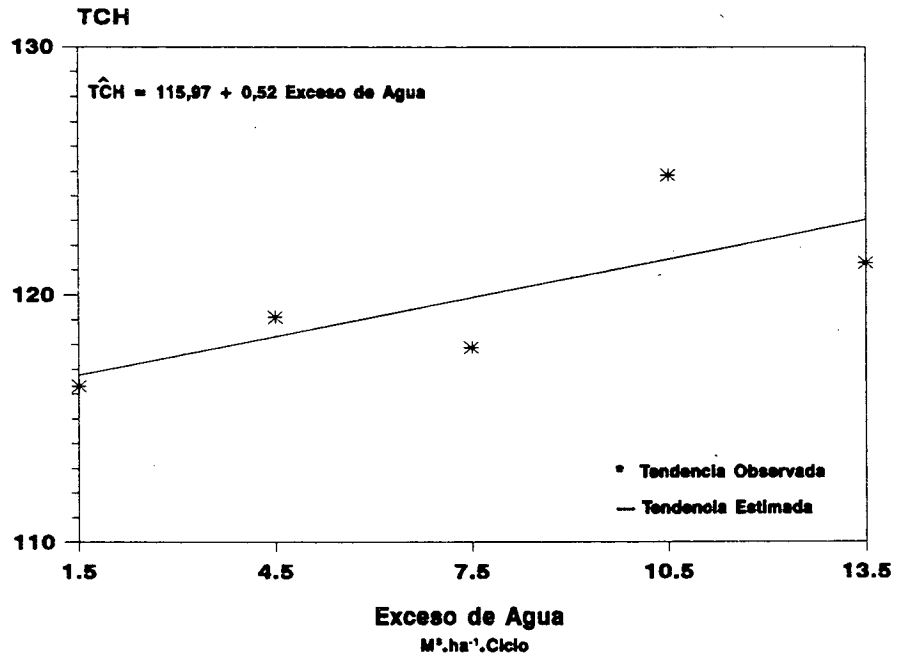


GRAFICO 1. Modelo para TCH en función del exceso de agua. Variedad CP 57-603.

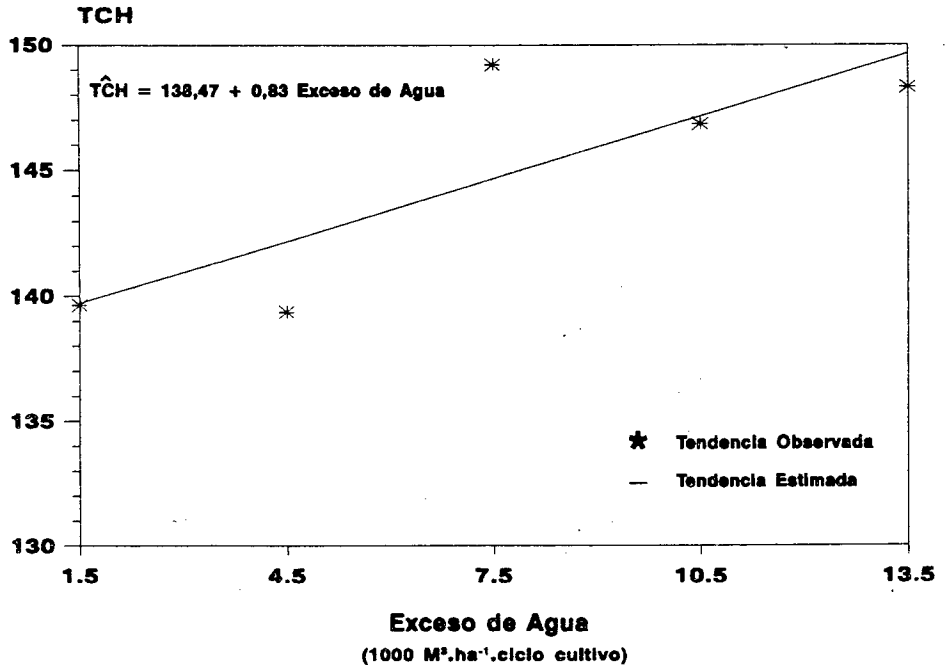


GRAFICO 2. Modelo para TCH en función del exceso de agua. Variedad MZC 74-275.

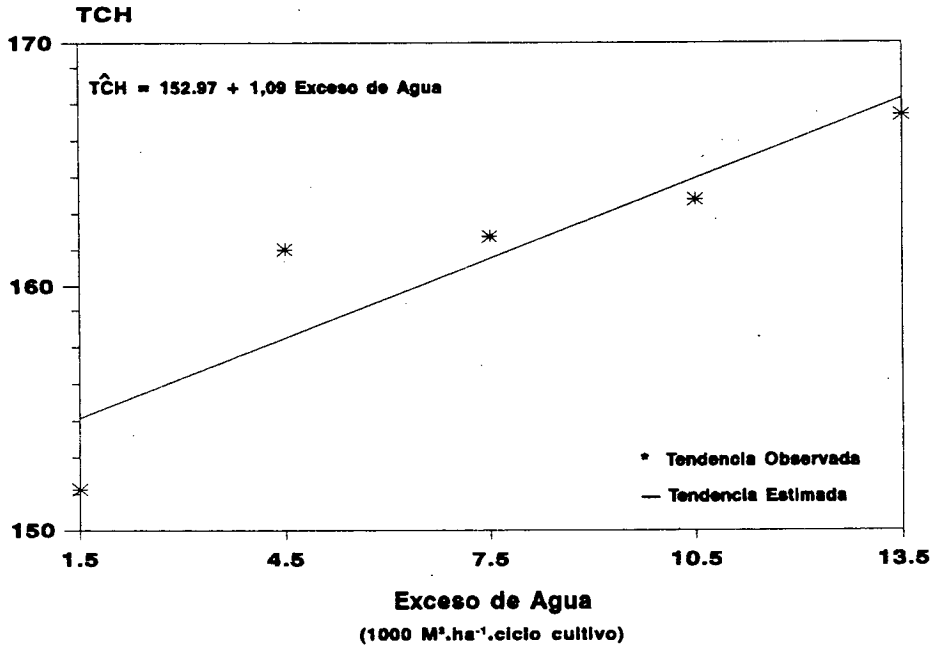


GRAFICO 3. Modelo para TCH en función del exceso de agua. Variedad PR 61-632.

Debido a que la edad de cosecha explica en cierto grado los resultados de producción, se calcularon estadísticas descriptivas para la edad de corte asociadas a cada grupo de exceso de agua; con el ánimo de ver si este factor presenta una tendencia con los incrementos del agua en el exceso. Un análisis de varianza detectó diferencias significativas entre variedades; no hubo suficiente evidencia para afirmar que estas diferencias dependen de los grupos de exceso de agua o viceversa, como tampoco hubo significancia para las otras interacciones. En el cuadro 3, se puede observar que la edad de corte es similar para todos los grupos de exceso de las variedades, por lo tanto no existe un efecto confundido de la edad de cortes sobre los incrementos de exceso de agua de las variedades consideradas.

CUADRO 3. Promedios de edad de cosecha de las variedades según grupos de exceso de agua.

Variedad	Grupo de Exceso de Agua (m^3)	Edad de Corte (mes)
CP 57-603	**1.** (0 -3000]	13,33
	2. (3000-6000]	12,97
	3. (6000-9000]	13,07
	4. (9000-12000]	13,35
	5. (12000-15000]	13,60
MZC 74-275	**1.** (0-3000]	13,49
	2. (3000-6000]	13,29
	3. (6000-9000]	13,43
	4. (9000-12000]	13,28
	5. (12000-15000]	13,41
PR 61-632	**1.** (0-3000]	14,85
	2. (3000-6000]	14,54
	3. (6000-9000]	14,73
	4. (9000-12000]	14,51
	5. (12000-15000]	15,59

2.4. EVALUACION ECONOMICA DE LOS EXCESOS DE AGUA

Acorde con las contribuciones marginales que obtienen cada una de las variedades ante los incrementos del agua en exceso; se valora en términos económicos la producción de caña adicional asociada a cada grupo de exceso y el costo del exceso de agua como tal. En el Cuadro 4 se puede observa que el valor de la producción adicional de caña por hectárea para todas las variedades y en todos niveles de exceso es menor que los costos incurridos por los cañicultores al regar una hectárea por encima de lo requerido; de tal manera que debidos al exceso de agua aplicado, este ingenio durante los tres años analizados presentó a nivel agregado unas pérdidas que ascienden a $1.361 millones de pesos corrientes de 1994, tomando como base para el cálculo el promedio de la cantidad de agua en exceso aplicada por hectárea durante el ciclo de cultivo; donde el promedio general de las pérdidas por hectárea son de $154.250 durante el ciclo de cultivo.

Dado que lo generalizado es encontrar excesos en la aplicación de agua, se toman las suertes que presentaron los menores niveles, con el ánimo de hacer unas cuentas conservadoras de las pérdidas incurridas por hectárea a nivel general y de las variedades; pérdidas que representan los posibles ahorros que se pueden lograr mejorando esta práctica agrícola. El 25% de las suertes menos intensivas en el consumo de agua en exceso tienen en promedio una pérdida $138.800 por hectárea, reduciéndose aproximadamente en 10% con respecto al promedio general y con base en este valor la sumatoria incurrida por el ingenio que no representó ninguna productividad adicional alcanzan un total de $1225 millones de pesos corrientes de 1994 en toda el área cosechada con las tres variedades.

CUADRO 4. Valor de la producción adicional de caña por hectárea y costos del exceso de agua aplicado por hectárea.

Variedad	Grupo de Exceso de Agua	Area Cose-chada (ha)	Valor Producción de Caña adicional ($*ha$_{-1}$)	Costo exceso de Agua ($*ha^{-1})	Pérdidas Económicas ($*ha^{-1})
CP 57-603	**1.** (0 -3000]	365	16.470	76.540	(60.070)
	2. (3000-6000]	1.393	41.545	175.060	(133.515)
	3. (6000-9000]	1.424	67.250	274.170	(206.920)
	4. (9000-12000]	473	91.500	336.080	(244.580)
	5. (12000-15000]	151	123.950	438.900	(314.950)
MZC 74-275	**1.** (0-3000]	247	25.660	70.330	(44.670)
	2. (3000-6000]	1.000	66.870	171.300	(104.430)
	3. (6000-9000]	1.351	107.400	258.900	(151.500)
	4. (9000-12000]	684	149.600	339.900	(190.300)
	5. (12000-15000]	249	190.730	423.700	(232.970)
PR 61-632	**1.** (0-3000]	235	42.230	95.520	(53.290)
	2. (3000-6000]	225	85.550	182.750	(97.200)
	3. (6000-9000]	543	142.530	276.630	(134.100)
	4. (9000-12000]	427	200.160	363.960	(163.800)
	5. (12000-15000]	58	261.400	432.870	(171.470)

Se observa que las variedades se concentran en una mayor proporción en los niveles de exceso de agua que van de 3000 m^3 a los 12000 m^3, siendo la CP 57-603 la que agrupa un poco más de la tercera parte en un sólo grupo de exceso.

2.4. CONCLUSIONES

- Análisis estadístico de este tipo de información permite detectar que labores o prácticas agrícolas generalizadas no contribuyen a incrementar las productividades de los cultivos en caña de azúcar.

- Ningún exceso de agua se justifica económicamente.

- La utilización de la teoría de modelos lineales permite cuantificar los efectos de factores ausales y sirven de referencia para retroalimentar y orientar el proceso investigativo.

PARTICIPATORY ON-FARM RESEARCH: STATISTICAL LITERATURE AND FUTURE POSSIBILITIES

by

Janet Riley

IACR - Rothamsted

Harpenden

AL5 2JQ

UK

A presentation prepared for the International Symposium on 'Statistics in Agriculture and Environmental Research' - Satellite Conference, Centro Internacional de Agricultura Tropical, Cali, Colombia. June 7-10, 1995.

Abstract

In response to requests from agricultural research workers in developing countries, the author initiated a project to establish the current availability of statistical literature for participatory on-farm research. A description is given in this talk of the range of material available and the different applications for which it is suitable. Indications are given of areas where statistical development is necessary to provide a comprehensive set of statistical material for both general and specific applications.

IACR receives grant-aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom.

Métodos Estadísticos en Biología Molecular

STATISTICAL MODELS FOR THE DETECTION OF GENES CONTROLLING QUANTITATIVE TRAIT LOCI EXPRESSION

Emilio A. Carbonell
IVIA, Apartado Oficial, 46113 Moncada (Valencia), Spain

Many important traits in plant breeding exhibit continuous variation (yield, maturity, biotic and abiotic stress tolerance, etc.). The genetic principles underlying their inheritance are basically the same as those affecting Mendelian or qualitative traits, but since the segregation of the genes concerned could be followed individually, new methods and concepts had to be developed. In this presentation, the task of searching for genes affecting quantitative, or continuous, traits will be considered. The immediate hope is the possibility of identifying specific portions of the genome involved in the variation of these traits (called QTLs) in order to enhance breeding programs. Moreover, the long term hope is finding the location of these genes to characterize and manipulate them to our advantage.

Therefore, we will try to discuss the different statistical models used for QTL analysis and the power of different approaches. QTL analysis have been approached using different statistical strategies depending on the number of markers involved in the analysis. Early studies considered the relationship between a marker and a QTL; later, models considered a pair of markers flanking the QTL and studied the association between the QTL and the interval defined by the flanking markers. More recently, statistical methodologies focus their attention to consider the whole linkage group considering all markers of the group as being associated with the QTL. Although any type of classification of statistical approaches is always incomplete and biased, we will try to describe the current statistical models organized according to the number of markers studied and by the basic statistical methodology being employed.

The practical implications of the approach will be discussed presenting a study on genotype by environment interactions for 15 traits in almonds investigated during 2-3 years by means of isozymatic markers.

MODELOS ESTADISTICOS PARA LA DETECCION DE GENES RESPONSABLES DE LA VARIACION DE CARACTERES CUANTITATIVOS

Emilio A. Carbonell

IVIA, Apartado Oficial, 46113 Moncada (Valencia), Spain

Muchos caracteres de interés agronómico muestran variación continua (rendimiento, precocidad, tolerancia a stress biótico y abiótico, etc.). Los principios genéticos que regulan su herencia son, básicamente, los mismos que afectan a los caracteres Mendelianos o cualitativos; sin embrago, dado que no se podía efectuar de forma individual la segregación de los genes responsables de su variación, ha sido necesario desarrollar nuevos métodos y conceptos. En esta conferencia, consideraremos el tema de la búsqueda de genes que afectan a los caracteres cuantitativos. El objetivo a corto plazo es la posibilidad de identificar regiones específicas del genoma involucradas en la variación de estos caracteres (o QTLs) para ser utilizados en programas de mejoramiento genético. A largo plazo, se espera llegar a localizar más finamente estos genes para caracterizarlos y manejarlos en nuestro beneficio.

Discutiremos los diferentes modelos estadísticos empleados en el análisis de QTLs y la potencia de los diferentes enfoques. Los análisis de QTLs se han efectuado siguiendo diferentes estrategias según el número de marcadores que se estudia simultáneamente. Los primeros estudios implicaban a un único marcador; posteriormente, se consideraron parejas de marcadores y se estudiaba la presencia de un QTL dentro del intervalo que definen. Más recientemente, los métodos estadísticos centran su atención en considerar todo el grupo de ligamiento, e incluso todo el genoma, y estudian las asociación de todos los marcadores simultáneamente. Aunque cualquier tipo de clasificación es incompleta y sesgada, describiremos los métodos existentes de acuerdo a número de marcadores que estudian y la técnica estadísitca que emplean.

Las implicaciones prácticas del enfoque del análisis de QTLs se discutirá presentando un estudio mediante marcadores isoenzimaticos sobre la interacción genotipo-medio en 15 caracteres de almendro evaluados durante 2-3 años

STATISTICAL MODELS FOR THE DETECTION OF GENES CONTROLLING QUANTITATIVE TRAIT LOCI EXPRESSION

Emilio A. Carbonell

IVIA, Apartado Oficial, 46113 Moncada (Valencia), Spain

1. INTRODUCTION AND BASIC CONCEPTS

The set of hereditary material transmitted from parents to offspring is known as the genome, and consists of molecules of DNA arranged in the chromosomes. The DNA itself is characterized by its nucleotid sequence. DNA sequences therefore have lengths measured in base pairs. A physical map is an ordering of features of interest along the chromosome in which the metric is the number of base pairs between features. However, we are concerned with genetic mapping where the metric itself is under genetic control.

Genetic map distances depend on the level of recombination expected between two points. An individual receives one copy of each heritable unit from each parent, but the combination of units at different locations the individual transmits to the next generation need not be one of the parental sets. Recombination may have taken place during the process of meiosis, and recombination between two elements on the same chromosome is more likely the further apart are those elements, with a limiting value of 50%. Although there is generally a monotonic relation between physical and recombinational distances, it is not a simple one. The distance over which one recombinational event is expected to occur depends on the region of the genome.

Detailed maps have been developed in model species like *Drosophila* many years ago and more recently in crop plants. These maps consisting of identifiable features on the genome at known locations or *markers*, can be used in the search for genes affecting traits of interest of discrete

nature like human diseases. These traits are the easiest to use because there is little ambiguity over which individuals had the disease.

Many important traits in plant breeding exhibit continuous variation (yield, maturity, biotic and abiotic stress tolerance, etc.) and therefore are named metric, polygenic, or quantitative traits (QTs). The genetic principles underlying their inheritance are basically the same as those affecting Mendelian or qualitative traits, but since the segregation of the genes concerned cannot be followed individually, new methods and concepts had to be developed. Why the segregation of genes involved in the variation of a QT in a population cannot be followed individually?. It is mainly due to two common well accepted aspects of the genetic control of QTs: a), "many" loci are involved, and b), the effect of the individual genes is "very small" compared to the environmental effects. Consequently, a branch of the science of Genetics concerned with quantitative traits grew up, the Biometrical or Quantitative Genetics.

In this discussion, the much more difficult task of searching for genes affecting quantitative, or continuous, traits will be considered. The fact that these traits are controlled by more than one gene, as well as by other non-genetic causes like the environment, further complicated the search. The immediate hope is that the possibility of identifying specific portions of the genome involved in the variation of this QTs in order to enhance breeding programs. Moreover, the long term hope is finding the location of these genes to characterize and manipulate them to our advantage. This basic problem is a very old one, In fact, Sax (1923) introduced for the first time the concept of using qualitative genes to locate genes of lesser effects controlling QTs. Thoday (1961) used single gene morphological markers to conduct detailed studies of QTs in *Drosophila melanogaster*. More systematic attempts to resolve QTs into their individual genetic components were initially limited by the scarceness of polymorphic qualitative markers that could cover large parts of the genome. These limitations were partly overcome by the use of isozymes, later by the restriction fragment length polymorphisms (RFLPs) which have provided a major source of markers with many desirable attributes and lately by RAPDs (random amplified polymorphic DNA), microsatellites and other markers at the DNA level. Thus, genetic or molecular markers (isozymatic loci, RFLPs and

RAPDs, etc.) linked to loci affecting a quantitative trait of interest can be used to follow their individual segregation and perform a kind of indirect selection named "marker assisted selection" to improve these traits more efficiently. The availability of detailed linkage maps of molecular markers makes it possible to dissect QTs into discrete factors, called QTLs (Quantitative Trait Loci) by Gelderman (1975); and, when a high enough number of markers are scored in a family segregating for a given polygenic trait, accurate estimates of genic effects and QTL locations can be obtained.

Molecular markers have been used to study many QTs in several species. Thus, great efforts have recently been focused on the construction of saturated linkage maps based on molecular markers to locate genes and QTLs affecting important traits by means of specific statistical methods. Four types of experimental designs are generally used in these studies: F_2's, backcrosses, recombinant inbred lines and doubled haploid lines. The first two designs are the most frequently used for gene mapping mainly due to the less time involved in annual crops, but the last two allow unlimited replications. Some annual crops have been extensively studied (tomato, corn, etc.) others, like perennial crops or forest trees due to their inherent complexities like long juvenile periods have been much less studied.

Therefore, in this presentation we will try to explain and discuss the different statistical models seed for QTL analysis, the power of different approaches and some aspects of specifical applications in almonds.

2. STATISTICAL METHODOLOGIES

Genetic mapping of QTLs rests on the simple idea that genetic markers that tend to be transmitted together with specific values of the trait are likely to be close to a gene affecting that trait. In other words, an association is sought between marker variants and trait values, with higher levels of association suggesting closer genetic map distances.

QTL analysis have been approached using different strategies depending on the number of markers involved in the analysis. Early studies considered the relationship between a marker and a QTL; later, models considered a pair of markers flanking the QTL and studied the association between the QTL and the interval defined by the flanking markers. More recently, statistical methodologies focus their attention to consider the whole linkage group considering all markers of the group as being associated with the QTL.

Although any type of classification of statistical approaches is always incomplete and biased, we will try to describe the current statistical models organized according to the number of markers studied and by the basic statistical methodology being employed.

2. 1. A marker at a time.

2. 1. 1. Linear models: Comparison of means and ANOVA

Traditional methods to estimate the association between a marker locus and a quantitative trait are based on a very simple and appealing approach: individuals are classified according to the genotype of the marker; when the means for the quantitative trait for those groups are not statistically different it indicates that the classification into groups was somehow arbitrary as far as the quantitative trait is concerned; therefore, the marker and the QTL are independent (i. e., no QTL is linked with the marker). Conversely, if means are statistically different, it is an indirect proof of the association between the grouping structure (the genetic marker) and the quantitative trait. So, a group of statistical methodologies are based on the comparison of means or ANOVA approach. Formally, the theory behind this intuitive approach is as follows:

Let us consider a population derived from a cross between two inbred lines. Denoting by M the marker locus and Q the locus involved in the expression of a quantitative trait. Consider two codominant alleles for the marker locus and that both loci are linked, r being their recombination fraction (2r would be the probability of crossing over during meiosis) independent of the sex of the parents from which the gametes were produced. Defining a and d after Falconer (1989) as the genotypic value of the homozygote and heterozygote genotypes of the Q locus, the phenotypic

value of an individual i could be expressed for a given QTL by the following linear model:

$$Y_i = \mu + aX_i + d(1-X_i^2) + \epsilon_i \qquad i = 1,.....,n \qquad (1)$$

where X is a dummy variable taking values of +1 (for the dominant homozygote), 0 (for the heterozygote) or -1 (for the recessive homozygote) in such a way that the expected genotypic values are:

$$\begin{aligned} \mu_{Q_1Q_1} &= \mu + a \quad (X = +1) \quad for \quad Q_1Q_1 \\ \mu_{Q_1Q_2} &= \mu + d \quad (X = 0) \quad for \quad Q_1Q_2 \\ \mu_{Q_2Q_2} &= \mu - a \quad (X = -1) \quad for \quad Q_2Q_2 \end{aligned} \qquad (2)$$

and ϵ_i is random variable that is assumed to be normally distributed with mean 0 and variance σ^2 and includes, among other things, the effect of other QTLs affecting the trait. This general model can be particularized for other mating designs.

Other authors use and alternative but equivalent model in terms of mean, additive and dominance effects:

$$\mu_{Q_iQ_j} = \mu + a_i + a_j + d_{ij} \qquad (3)$$

F_2 and backcross populations have been extensively studied by many authors. Most of them used the ANOVA approach and presented the frequencies of the QTL genotypes (Table 1), and the expected values of the means and variances (Table 2) of the quantitative trait within each marker genotypic class. If both loci segregate independently (r = 0.5), the means of the quantitative trait for each genotypic marker class are both equal; therefore, a way to test if r = 0.5 is by testing the null hypothesis of equal class means by a t-test using non-pooled variances because they are different due to dominance and/or linkage between the marker locus and the QTL; hence, standard

ANOVA procedures should be avoided in this case. Variances will be equal in the case of no association (which is precisely what we want to test) or under no dominance of the trait. Moreover, even though the t-test is quite powerful for departures from non normality, it should be recognized from the structure of table 2 that within each marker class we have a mixture of normals where proportions are functions of the recombiantion fraction.

For the F_2 population, the contrasts of interest to estimate the genetic effects are:

$$\mu_{M_1M_1} - \mu_{M_2M_2} = 2a(1 - 2r)$$
$$2\mu_{M_1M_2} - (\mu_{M_1M_1} + \mu_{M_2M_2}) = 2d(1 - 2r)^2$$

(4)

Therefore, the estimates of the genetic effects are biassed due to the confounding effect of the recombiantion fraction. Since the quantity in parenthesis is always equal or lower than one, the bias in the estimate of the dominant effect is larger than that of the additive effect. Unless the QTL shows a complete dominance (a = d), there is no way to distinguish between a QTL tightly linked having a small genetic effect from another with a large effect but loosely linked. Hence, this type of approach could be considered as preliminary or exploratory to simply detect the presence of a QTL. Therefore, to solve the mixture problem and to obtain unbiased estimates of both the magnitude of the gene effects and the recombination fraction are needed, alternative statistical methods should be used.

2. 1. 2. Maximum likelihood estimation

To have a separate estimate of the recombination fraction and the biometrical parameters of the QTL, alternative methods have been developed. Maximum likelihood has been shown to be more powerful than moments method for certain problems of estimation; however, it requires that the type of distribution be known. Many studies apply the central limit theorem and assume an underlying normal distribution of the trait unless there is reason to assume otherwise.

For a F_2 population, the method is as follows: Let Y_i, Y_j and Y_k denote trait values for three random individuals belonging to the maker genotypic classes M_1M_1, M_1M_2 and M_2M_2, respectively. The likelihood of the entire population can be written as:

$$L=\prod_{i=1}^{n_1} f(Y_i)\prod_{j=1}^{n_2} f(Y_j)\prod_{K=1}^{N_3} f(Y_K) \qquad (5)$$

where L is the likelihood function; $f(Y_i)$ is the density function of Y_i; and n_i is the total number of individuals in the ith marker class. In turn, the density functions are:

$$\begin{aligned} f(Y_i) &= (1-r)^2 f(Q_1Q_1) + 2r(1-r) f(Q_1Q_2) + r^2 f(Q_2Q_2) \\ f(Y_j) &= r(1-r) f(Q_1Q_1) + (1-2r+2r^2) f(Q_1Q_2) + r(1-r) f(Q_2Q_2) \\ (Y_k) &= (1-r)^2 f(Q_2Q_2) + 2r(1-r) f(Q_1Q_2) + r^2 f(Q_1Q_1) \end{aligned} \qquad (6)$$

That is, each density function is expressed as the sum of three disjoint events: since the linkage between marker and QTL is incomplete, the individuals within each marker class are a mixture of the three genotypes at the QTL. In turn, the probability of each event is a set of conditional probabilities: the product of the probability of having a particular genotype at the Q locus given the genotype of the marker (calculated from table 1), times the density of Y_i, conditional on the fact that the individual has that genotype at the QTL (f(QQ)). In order to obtain the form of the latter density, the underlying distribution of the genotypes at the QTL in the entire population (not within a marker genotypic class) is assumed to be normal. The variances include not only environmental variation but also genetic variation at other loci (QTL) affecting the quantitative trait.

For, say, $f(Q_1Q_1)$ the conditional density will have the form:

$$f(Q_1Q_1) = (2\pi\sigma_{Q_1Q_1})^{-1/2} \exp[-\frac{(Y_i - \mu_{Q_1Q_1})^2}{2\sigma^2_{Q_1Q_1}}]$$

(7)

In order to obtain the estimates of the recombination fraction and the means and variances, partial differentials of the log of the likelihood function L for all seven parameters are set to zero and solving the resultant equations. If an analytical solution is not readily evident, iterative solutions are available, but they tend to be cumbersome and time consuming and not necessarily provide solutions.

The problem with the approach is the large amount of computational resources that are required. Furthermore, because of the need to search the likelihood surface for each parameter separately and ignorance of the internal mathematical relationships among the parameters to be estimated, this method could not guarantee that the estimates obtained were in fact maximum likelihood estimates.

2. 2. Pair of markers

The basic idea about using a pair of markers flanking the putative QTL was put forward by Lander & Botsein (1986; 1989). Once a map of markers is available, we test whether a QTL lies in an interval of known size between two adjacent markers.

If the putative QTL is located within the interval defined by the two marker locus A and B with known recombination fraction between them equal to p, we denote by r and r' the recombination fractions to the left and right markers. If the distance between the two markers is very small, one may assume that the frequency of double cross-over is negligible; then $p = r + r'$. If not, $p = r + r' - 2rr'\delta$, where δ is the coefficient of coincidence. Under no interference, $\delta = 1$, hence using the Haldane function $p = r + r' - 2rr'$

This approach combines information on all markers into one likelihood map. ~~The~~ integration of all

markers makes this method very appealing, giving detailed insight into the part of the genome being investigated. The testing for the presence of a QTL (null hypothesis of $H_0: a=0$ corresponding to the hypothesis that no QTL is linked), is given by the significance of the linear model as shown in equation (1) using the Fisher-Snedecor F statistic.

When the QTL is located exactly at one of the two markers defining the interval, the problem is easily solved by standard regression procedures because in this case, the genotype of the QTL is the same as the marker; so, the values of X_i in equation (1) are known. However, if the QTL is located somewhere between the flanking markers (which is the normal situation), these values cannot be directly obtained but only their probability distributions. In this case, the likelihood of the parameters $\theta=(\mu,a,\sigma^2)$ based on the observations $Y=(Y_1,Y_2,.....Y_n)$ is given for a F_2 population by:

$$L(\theta|Y) = \prod_{i}^{n} [G_i(-1)\, z(y_i| -1;\theta) + G_i(0)\, z(y_i|0;\theta) + G_i(+1)\, z(y_i| +1;\theta)] \qquad (8)$$

Where $G_i(X)$ represents the probability distributions of X over the values of -1, 0, and +1 given the genotypes of the flanking markers for individual i. This probability is a function of p, r and r'; and, given that p is known and that r' can be expressed in terms of p and r assuming the Haldane function ($p = r + r' - 2rr'$), this probability has only r as unknown. Therefore, the recombination fraction between the QTL and the left marker of the interval, r, is included in the likelihood through $G_i(X)$. Hence, for each value of r, a different likelihood will be obtained. Values of the probabilities for all combinations of genotypes of the markers are shown in Table 3. This probability will be calculated for each individual i according to the specific genotype of its markers. The function

$$z(y_i|x_i;\theta) = (2\pi\sigma^2)^{-1/2} \exp[-\frac{(y_i - \mu - a x_i - d(1 - x_i^2))^2}{2\sigma^2}]$$

(9)

denotes the conditional density of Y_i given $X_i = x_i$ (-1, 0, or +1) assuming normal distribution of the trait. In order to obtain the maximum likelihood estimates of θ, the function $L(\theta|r)$ may be maximized through choices of θ. This may be accomplished by using the EM algorithm (Dempster et al., 1977) for a fixed r.

The E and M steps should be repeated for convergence of the estimates of θ. Once the parameters are obtained, the most likely position may be estimated by the maximum values of the F statistic, through all possible values of r varying along the intervals. The F values could be plotted against r; then, the most likely position of the QTL is given by the value of r, say r_M, which gives the

maximum of the curve provided it is higher than a given significant threshold. Probability statements when testing the hypotheses that certain effects are zero (such as a = 0 and/or d = 0) can be established.

The effect of two linked QTLs on the estimation of the recombination fraction and the genetic effects using the above model was studied with simulated data. Results from 100 simulations consistently gave bimodal distributions for the higher heritability of the trait of 0.50 as shown in Figure 1 (curve A for doubled haploids and curve B for backcrosses). For doubled haploids and heritability of 0.08, few cases showed a highly significant unimodal distribution (curve C) like the "ghost" QTL observed by Martinez & Curnow (1992), or a multimodal or plateaued curve (curve D) indicating that QTLs could be found anywhere within a long range of values of the recombination fraction with no clear location of the maximum of the likelihood function. More frequently, a bimodal curve (curve E), like in the high heritability case, indicated the presence of

two QTLs; however, the predicted locations were biased in the sense that the leftmost QTL was shifted to the right and the leftmost one to the left (see curve F for the comparison of predictions).

Therefore, implicit in the above methods is the fact that a single QTL is investigated at a time and that it segregates independently from the remaining QTLs contributing to the quantitative trait. The full model is actually a summation of the a_j and d_j effects over the total number of QTLs (j = 1, 2,....q) so that when a particular QTL is investigated, it is assumed that its effect is not influenced by the rest of the QTLs. Therefore, it is not intended to resolve clusters of QTLs but only those that are not too close to make tenable the assumption of independent action on the measured quantitative trait.

2. 3. All markers simultaneously.

Multiple regression techniques have been suggested to identify multiple QTLs linked to several markers.

The basic procedure consists of assigning a coded value to each marker genotype and use stepwise regression procedures to identify which markers are associated with the variation of the quantitative trait The multiple regression model will contain quadratic terms to account for dominance at the QTLs. The problems with this approach are those of the stepwise selection of variables: several marked segments are usually available in the same chromosome for the analysis of the QTL effects; these segments do not likely segregate independently. Since stepwise regression procedures incorporate the variables into the model in a sequential manner, one by one, the estimates of the effects of those QTLs which are already in the model may be biased by other linked and unlinked QTL which may or may not be yet in the model. Some QTLs already in the model may prevent the entrance of a new QTL in the model; therefore, some QTL may cover the presence of another QTL that will remain undetected by the statistical methodology. Moreover, the method may produce very different results depending on the path of selection followed by the computer programs; i. e., if two variables have very similar F values to be in or out of the equation and are highly correlated, the final equation may depend on which of the two variables is chosen

to enter in the equation.

Considering the genome as a whole, Rodolph & Lefort (1993) proposed a method based on the linear model that can be used in a very general genetic context. The purpose is to detect regions displaying QTL effects rather than to detect isolated QTLs and estimate their effects. Their linear model is actually an extension of the basic model in equation (1) summing over the set of m markers, that is:

$$E(Y_i) = \mu + \sum_{j}^{m} \begin{cases} + \alpha_j - \delta_j & for\ M_i(j) = M_1M_1 \\ + \delta_j & for\ M_i(j) = M_1M_2 \\ \alpha_j - \delta_j & for\ M_i(j) = M_2M_2 \end{cases} \qquad (10) \qquad \text{for an } F_2$$

where $M_i(j)$ represents the value taken by the jth marker (j = 1,...m) on individual i. Parameters α_j and β_j are statistical effects associated with the markers and are related to the additive and dominance effects.

The major advantage of the methods based on all markers simultaneously is the possibility of making global tests, recovering the whole or a part of a chromosome. In doing so, the existence of a set of linked QTLs can be detected. Conversely, the main characteristic of the interval mapping methods is their sequential nature. This is the reason why they are not well suited for the detection of closely linked QTLs, but they are more powerful in the case of an isolated QTL. Hence, a compromise should be reached using both approaches. In fact, recently, a joint approach has been proposed that makes the combined use of the conventional interval mapping with regression methods in order to detect multiple QTLs. It was proposed by Jansen (1993) suggesting to fit one QTL at a time in a given interval and simultaneously using some of the markers as cofactors to eliminate the effect of additional QTLs. Similar idea was put forward by Zeng (1993, 1994) and expanded by Jansen and Stam (1994) in a very general method. The method exploits two features,

the use of additional parental and F_1 data, which fixes the joint QTL effects and the environmental error, and the use of markers as cofactors which reduces the genetic background noise. By using the EM algorithm, missing values of any kind (markers or quantitative traits) are easily estimated making full use of the data.

Most methods are based on segregating populations derived from crosses of inbred lines. For many species this is not practicable, but crosses can be made between outbred lines. Haley et al. (1994) used least-squares methods to regress trait phenotypes onto additive and dominance effects of putative QTLs in marker intervals. The work is for the situation of crosses between outbred lines in which the trait loci are segregating but in which the markers used are fixed for alternative alleles.

3. SAMPLE SIZE AND POWER STUDIES

The accuracy of any mapping procedure depends not only on the ability of the statistical method to determine the location and to estimate the genetic effect of the QTLs. Other factors have a very important influence on this accuracy as well: the heritability of the trait, the contribution of each QTL to the total genotypic variance, the number of QTLs, their interactions, their distribution over the genome, their distance to the markers, the statistical distribution of the random non-genetic factors, the type of segregating population studied, its size, the genome size, the number of marker loci employed, etc. Given the large number of factors, authors restrict themselves to partial studies investigating some of these factors at a time.

For the single marker approach, Soller et al. (1976) gave approximate formula to get the number of individuals per marker class required to detect a given difference between the values of the quantitative trait in each marker class. They assumed that the QTL was located exactly on the marker (which is quite unlikely to happen); for F_2 they wrongly assumed constant variance of the quantitative trait within the marker classes. According to their calculations they concluded that a few thousand offspring would be enough to detect close linkages involving QTLs contributing about 1% of the total phenotypic variance of the trait in the F_2. Relaxing both assumptions Asíns & Carbonell (1988) gave the following formula for the number of individuals per homozygote

genotype in a F_2 population in order to detect the QTL at the 5% level and a probability of error Type II of 10%:

$$N \geq \frac{10.5\left[r(1-r)+k^2r(1-3r+4r^2-2r^3)-\frac{1}{4}(1+\frac{1}{2}k^2)(1-\frac{1}{h^2})\right]}{(1-2r)^2} \qquad (11)$$

where k is the degree of dominance ($k=\frac{d}{a}$), and h^2 is the heritability of the trait. It is clear from the formula that the number of offspring is a function of the unknown recombination fraction; hence the maximum allowed recombination fraction should be stated in order to obtain the sample size. As a practical guide, Cowen (1988) indicated that the smallest additive effect *a* detectable is in the range of 1/2 to 1/12 the size of the corresponding LSD at the 5% for the quantitative trait. He used 100 recombinant inbreds or doubled haploid lines and 6 replicates. With S_1 lines, the within-population dominance effects *d* can be also detected. The smallest *d* detectable is 3-4 times the above size.

The comparison between the maximum likelihood approach and the comparison of means seems to be influenced by the poor convergence during maximum likelihood estimation when the effects of the QTL are small and for recombination fractions approximately 0.5.

For the interval mapping approach, Carbonell et al. (1993) used 100 replicated sets of 250 simulated individuals to investigate the power of the interval mapping to detect and estimate the genetic effects of several QTLs. The simulated individuals had eight linkage groups with six markers each separated by 20 cM distance. Six unlinked QTLs were involved in the expression of the quantitative trait, with different gene action under the assumption of dominance in some QTLs or a completely additive model. They concluded that doubled haploid populations could be used with smaller sample sizes because they show much higher power than backcrosses. Moreover, more accurate estimates of the location of the QTL and with less variance were obtained. This

result could be expected from the fact that the model uses more spread-out values (+1 and -1) than backcrosses (0 and -1), and that the absolute difference between the means of both homozygous genotypes is larger than that between the heterozygous and the recessive homozygous genotypes. When dominance is present backcrosses not only give biased estimation of the effects, because additive and dominant effects are completely confounded in this mating design, but also some QTL could not be detected due to similar the genotypic values of the homozygote and the heterozygote at those loci. That could be the case for complete dominance when the F_1 population is backcrossed with a "high" producing parental line and also when, independent of the direction of the backcross, the recurrent parental line had the "high" allele at some QTL. For doubled haploids, the power of detecting a given QTL is clearly related to its relative contribution to the heritability of the character; i. e., proportional of to the square of its additive value a_j^2. The higher its contribution, the higher the probability of being detected. Even QTLs having small effects were identified by the model; the power of the test was about 90% for heritabilities of the QTL as low as 5%. To obtain similar power for backcrosses, the heritability attributable to the individual QTL should be around 14%. For a given type of gene action, it is difficult to compare with F_2 populations results (Carbonell et al., 1992) because two different criteria were used for this population but it seems that doubled haploids have a similar power than F_2. In fact, similar values of 5% for the heritabilities of the QTL were obtained by van Ooijen (1992) for this type of population. However, if dominance is present, doubled haploids will only detect the additive component of a particular QTL. This fact could be of an extreme importance for QTLs showing overdominant effects or to design the most efficient breeding strategy in order to exploit the non-additive variation hidden in some QTLs. The major technical advantage for doubled diploids, independent of any effect of replication on the required number of offspring, lies in the fact that the lines can be reproduced independently and continuously evaluated with respect to additional quantitative traits and markers with all the information being cumulative (Burr et al., 1988). If the effect of replication is taken into account, replicated progenies can bring about a major reduction in the number of individuals that need to be scored. Reductions are the greatest when heritability of the trait is low, but are much less when heritability is moderate to high under the assumption of codominance at all QTL (Soller & Beckmann, 1990).

Regarding the statistical approach it is interesting to compare the classical "single marker" methodology with the interval mapping. The methodologies based on the study of all markers simultaneously are much more specifically designed for the case of multiple linked QTLs.

Lander & Botsein (1989) concluded that even in the worst case where the QTL is located in the middle of the interval, interval mapping decreases the number of offspring by a factor of (1 - p) as compared with the single marker approach; so for maps having markers at distances of 10, 20, 30 and 40 cM, the savings are about 9%, 16%, 23% and 28%, respectively (assuming the Haldane mapping function). However, Knott & Haley (1992) compared both maximum likelihood procedures for simulated F_2 data and found that the use of flanking markers only provides substantially more power for QTL detection when markers are widely spaced and the QTL effects are large; however, interval mapping has the additional advantage of correctly positioning the QTL and estimates its effect on the average. The estimates obtained using single markers are not close to the simulated values with both the distance from the nearest marker and the additive effect overestimated. In this comparison, the authors have ignored the problem of the level of significance of the comparison and did not give the exact distributions of the test statistics they used. Similar results were obtained by Rebai et al. (1994) when analytically compared the power of the maximum likelihood using interval mapping and the ANOVA approach for backcross populations. They showed that the ANOVA test is less powerful for all non null values of the distance between the flanking markers; however, for small intervals (distances from 0 to 30 cM) the difference in power between the two tests do not exceed 5%. For large intervals and quite large QTL effects the advantage of interval mapping is 10% to 30%. For simplicity in the analytical derivations, their approach assumed absence of double crossing over; therefore, the results for large intervals would be approximate and conclusion should be taken with caution. For small effects of the QTL (less than 1% of the phenotypic variance) powers of the tests are close to the significance level and could not be compared.

Single marker methodologies based on ANOVA and comparison of means are much more simple to implement using standard statistical computer packages but they can be only used to the

detection of several unlinked QTLs. They can be used as an exploratory phase of the study where we are simply interested in the detection of association. If an estimate of the localization of the QTL and an unbiased estimation of effect is required, interval mapping should be used although previous mapping of the markers is needed. If precise estimation of the recombination fraction between the flanking markers is not available, Knott & Haley (1992) found that the use of an incorrect value for the recombination frequency between the flanking markers, does not affect the ability to map the QTL as long as the order of the markers along the chromosome is correct. On the other hand, when a QTL is located in the distal part of the linkage group not flanked by markers, the maximum of the LOD score is likely to arise in the interval near to the QTL, or even at the position of the marker; that will result in a mislocalization of the QTL and an under-estimation of its effect. Therefore, as Rodolphe & Lefort (1993) conclude, the comparison between results obtained from different statistical methods can give an interesting insight especially on the number of QTLs; if the purpose is a precise dissection of the genetic determinism of the trait under study, this cannot be achieved with only one experiment; the search for QTLs should be an iterative process. The multimarker model enables the experimenter to focus on the interesting segments detected. New experiments must be made, with more markers on these segments and more individuals in order to get more recombinant genotypes between these closely linked markers. These new experiments can be analyzed using different models and methods adapted to specific situations and purposes. Then, simultaneous use of several methods should lead to better efficiency in QTL technology.

4. PERSPECTIVES AND FURTHER REMARKS

In the preceding sections, it has been shown that there are still some unsolved matters that deserve further research. They are mainly related to the presence of closely linked and interacting QTLs, the development of efficient ways to detect QTL showing small effects, unveiling QTLs masked by other QTLs having very large effects, the presence of epistatic effects, methods not only directed towards increasing the power of the designs but also to decrease the probability of false positives, better knowledge of the distributional properties of the test statistics being used, models

for non-normally distributed traits, models for populations of allogamous plants, etc.

There has been a fast development of techniques in the areas of genetic markers and pollen-derived homozygous plants. Both technologies assisted by specific statistical methods provide powerful means for QTL identification. Carbonell et al. (1993) concluded for interval mapping that doubled haploid populations will allow to conduct experiments with less number of individuals because, similar to F_2 populations, they show much higher power than backcrosses and the estimates are more accurate. However, if dominance or overdominance play an important role at some QTLs, this information will not be unveiled using the doubled haploid design. As it has been discussed previously there are still left several situations where location of QTLs may fail or even providing false positives; therefore, further research efforts have to be focused on the statistical methods. Sometimes, it may be a problem not related to the statistical model itself but to the experimental design used. If the purpose is a precise dissection of the genetic determinism of the trait under study, we do not believe it can be achieved with only one experiment or methodology; QTL analysis has to be an iterative process.

In addition to QTL location and estimation of gene effects, important aspects of QTL analysis have been recently studied. Therefore, there is a considerable body of knowledge that should enable us to deeply investigate other well known and important phenomena related to the nature of quantitative variation such as: pleiotropy, gene interactions, heterosis, transgressive segregations, and genotype by environment interaction. Very recently, several authors have focused studies towards these subjects (de Vicente & Tanksley, 1993; Bretó et al., 1994; Asíns et al., 1994) and we will discuss now some of these aspects.

5. APPLICATION TO STUDY THE INTERACTION GENOTYPE BY ENVIRONMENT IN ALMONDS (ASÍNS ET AL., 1994)

The genotype by environment interaction (G x E) is the differential genotypic expression across environments. This interaction reduces the association between genotypic and phenotypic values and may cause selections from one environment to perform poorly in another, thereby forcing plant breeders to examine genotypic adaptation. In addition, the sampling problems associated with yearly variation suggest a necessary testing for many crop cycles. To save time, breeders opt to substitute temporal with spatial variation, assuming that testing over a wide geographic range can ensure a parallel degree of temporal buffering capacity in the germplasm.

QTL analysis based on genetic markers could be used as an approach to study G x E, and even more, to gain stability by designing a marker assisted selection scheme that takes into account all the QTLs involved in the trait for the different environments.

The objective of the investigation was to study the effect of years on QTL detection for 15 traits. The plant material consisted of 123 seedlings established in 1896 that had been derived by hand pollination of emasculated flowers of the Spanish almond variety "Ramillete" with pollen of the Italian variety "Tuono". Fifteen traits were evaluated in 1989 and 1990, and six of them also in 1991 according to Table 4.

The association of isozymatic markers and traits recorded as categorical variables was studied by chi-square contingency tables and by one-way ANOVA and t contrast for the remaining variables. These results are shown in Tables 5 and 6.

At least 22 putative QTLs were detected. Only 3 (or 5 at 10% significance) of them behave somehow homogeneously through the years, none of them were associated with traits recorded as categorical variables. Three factors may be involved in this lack of stability: 1), the test statistic used to detect association being dependent on the character definition of variables; 2), the contribution of the QTL to the total genetic variability of the character; and 3), a differential gene

expression depending on the year, i. e. due to G x E. The first two facto factors are close related. As it has been shown (Carbonell et al,. 1992), the power to detect a given QTL is related to its contribution to the heritability of the trait. Thus, with respect to moderate heritabilities (Dicenta et al., 1993a, b), if specific the contribution of the QTL is low, it may remain undetected. That could explain the cases like rugosity of the seed, RS, (tables 4 and 5) and percentage of the kernel, PK, (tables 4 and 6). If heritability is low and the statistical test is not very powerful, it is a matter of chance that the association is detected or not. This could have happen in case like the QTL detected for the color of the tegument, CT. Most cases in which there is a lack of stability in QTL detection involve traits whose heritability estimates change drastically from one year to another like yield intensity, YI, and number of double kernels, ND, and duration of flowering, DF, and/or whose correlation coefficients between years are low, like YI (from -0.14 to 0.28), kernel weight, KW, (0.36), and DF (from 0.00 to 0.24). This suggests that there are important differences in the number and/or the relative contributions of the genes controlling the quantitative trait that are dependent on the year (G x E).

It is noteworthy that the variation for three traits related to yield (YI, KW and PK) in 1990 was associated to segregation at Prx_c-2. Temperature records showed that the winter of that year (December 1989 to February 1990) was the warmest, with no under 0°C temperatures. Therefore it is not due to resistance to adverse conditions but a matter of genotype by environment interaction playing an important role in the determination of the phenotypic value of the plant.

Most associations found in 1989 were also found in 1990; however, none of the associations found in 1991 were found in the other years. The coldest winter was that from December 1990 to February 1991 with temperatures of -3°C. Similarly, the lowest values of the coefficient of correlation per trait among years were between 1991 and 1989 or 1990 (data not shown). Winter temperature regimes affect flowering times and the duration of flowering. The way these traits become affected must be a change in gene expression: one or several QTLs linked to Got-2 are involved in IF, MF and FF in 1989 and 1990, while other(s) linked to Prx_c-2 are specially relevant for those traits in 1991. This difference might be just a difference in the level of gene expression

at the QTLs involved in such a way that the final result is a change in their genic effects (losing or gaining importance in their contribution to the genotypic value of the trait).

Given that a common marker was associated with several traits, correlation coefficients among those traits per year were studied. High correlations (about 0.7) were found among IF, MF and FF for the three years; DF and FF for 1989 (0.71), and DF and IF for 1991 (-0.73). On the basis of these data and the simultaneous change of association of IF, MF and FF with other marker loci being dependent on the year, these traits must involve many QTL in common (QTLs with pleiotropic effects).

Hence, it has been shown that it is possible, not only to measure but also uncover the differential gene expression involved in G x E using the methodologies developed for QTL analysis and it is an essential step to establish the marker assisted selection scheme and to dissect the quantitative trait itself.

REFERENCES

Asíns, M. J., P. Mestre, J. E. Garcia, F. Dicenta & E. A. Carbonell, 1994. Genotype x environment interaction in QTL analysis of an intervarietal almond cross by means of genetic markers. Theor. Appl. Genet., 89: 358-364

Asíns, M. J. & E. A. Carbonell, 1988. Detection of linkage between restriction fragment length polymorphism markers and quantitative traits. Theor. Appl. Genet. 76: 623-626.

Bretó, M. P., M. J. Asíns & E. A. Carbonell, 1994. Salt tolerance in *Lycopersicon* species. III. Detection of quantitative trait loci by means of molecular markers. Theor. Appl. Genet. 88: 395-401.

Carbonell, E. A., M. J. Asíns, M. Baselga, E. Balansard & T. M. Gerig, 1993. Power studies in the estimation of genetic parameters and the localization of quantitative trait loci for backcross and doubled haploid population. Theor. Appl. Genet. 86: 411-416.

Carbonell, E. A., T. M. Gerig, E. Balansard & M. J. Asíns, 1992. Interval mapping in the analysis of nonadditive quantitative trait loci. Biometrics 48: 305-315.

Cowen, N. M., 1988. The use of replicated progenies in marker based mapping of QTLs. Theor. Appl. Genet. 75: 857-862.

Dempster, A. P., N. M. Laird & D. B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc., Series B 39: 1-38.

Dicenta, F., J. E. García & E. A. Carbonell. 1993. Heritability of flowering, productivity and maturity in almond. J. Hort. Sci. 68: 113-120.

Dicenta, F., J. E. García & E. A. Carbonell. 1993. Heritability of fruit characters in almond. J. Hort. Sci. 68: 121-126.

Falconer, D. S., 1989. Introduction to Quantitative Genetics. Longmans, London.

Gelderman, H., 1975. Investigation on inheritance of quantitative characters in animals by gene markers. 1. Methods. Theor. Appl. Genet. 46: 319-330.

Haley, C. S., S. A. Knott & J. M. Elsen, 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136: 1195-1207

Jansen, R. C., 1993. Interval mapping of multiple quantitative trait loci. Genetics 135: 205-211

Jansen, R. C. & P. Stam, 1994. High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136: 1447-1455

Knott, S. A. & C. S. Haley, 1992. Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genet. Res. Camb. 60: 139-151.

Lander, E. S. & D. Botsein, 1986. Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc. Natl. Acad. Sci. USA 83: 7353-7357.

Lander, E. S. & D. Botsein, 1989. Mapping mendelian factors underlying quantitative traits using RFLPs linkage maps. Genetics 121: 185-199.

Martinez, O. & R. N. Curnow, 1992. Estimating the location and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. 85: 480-485.

van Ooijen, J. W., 1992. Accuracy of mapping quantitative trait loci in autogamous species. Theor. Appl. Genet. 84: 803-811.

Rebai, A., B. Goffinet & B. Mangin, 1994. Comparing power of different methods for QTL detection. Biometrics (in press).

Rodolphe, F. & M. Lefort, 1993. A multi-marker model for detecting chromosomal segments displaying QTL activity. Genetics 134: 1277-1288.

Sax, K., 1923. The association of size differences with seed coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics 8: 552-560.

Soller, M. & J. S. Beckmann, 1990. Marker-based mapping of quantitative trait loci using replicated progenies. Theor. Appl. Genet. 80: 205-208.

Soller, M., T. Brody & A. Genizi, 1976. On the power of experimental designs for the detection of linkage between marker loci an quantitative loci in crosses between inbred lines. Theor. Appl. Genet. 47: 35-39.

Thoday, J. M., 1961. Location of polygenes. Nature 191: 368-370.

de Vicente, M. C.& S. D. Tanksley, 1993. QTL analysis of transgressive segregation in an interspecific tomato cross. Genetics 134: 585-596.

Zeng, Z. B., 1993. Theoretical basis of precision mapping of quantitative trait loci. Proc. Natl. Acad. Sci. USA 90: 10972-10976

Zeng, Z. B., 1994. Precision mapping of quantitative trait loci. Genetics 136: 1457-1468

Table 1. Frequencies of QTL genotypes within each genotypic class of the marker for an F_2 population.

Markers	Q_1Q_1	Q_1Q_2	Q_2Q_2
M_1M_1	$(1-r)^2$	$2r(1-r)$	r^2
M_1M_2	$r(1-r)$	$(1-2r+2r^2)$	$r(1-r)$
M_2M_2	r^2	$2r(1-r)$	$(1-r)^2$

Table 2. Means and variances of the QTL within each genotypic marker class for F_2 populations.

Marker genotype	Means	Variances [a]
M_1M_1	$\mu+a(1-2r)+2dr(1-r)$	$2a^2r(1-r)+2d^2ru-4adr(1-3r+2r^2)$
M_1M_2	$\mu+d(1-2r+2r^2)$	$2a^2r(1-r)+2d^2ru$
M_2M_2	$\mu-a(1-2r)+2dr(1-r)$	$2a^2r(1-r)+2d^2ru+4adr(1-3r+2r^2)$

(a) $u=1-3r+4r^2-2r^3$

Table 3. Marginal marker genotypic frequencies and expected and conditional frequencies of the two types of QTL genotypes given the genotypes of the flanking markers for a doubled haploid population.

		Conditional frequencies			
		Under Haldane's		Under no double	
Marker	Marginal	Q_1Q_1	Q_2Q_2	Q_1Q_1	Q_2Q_2
A_1B_1/A_1B_1	(1 - p)/2	(1 - s)	s	1	0
A_1B_2/A_1B_2	p/2	(1 - t)	t	r'/p	r/p
A_2B_1/A_2B_1	p/2	t	(1 - t)	r/p	r'/p
A_2B_2/A_2B_2	(1 - p)/2	s	(1 - s)	0	1

p is the recombination fraction between the markers, s = rr'//1 - p), t = r(1 - r')/p, (assuming Haldane's function); and r and r' are the distance of the putative QTL to the left (A) and right (B) marker, respectively

Table 4.- Coding, heritabilities and mean values of the traits (taken from Dicenta and Garcia 1993)

			Mean values		h^2		
Traits	Code	Years	Ramillete	Tuono	1989	1990	1991
Initial Flowering	IF	3	39.3	54.7	0.95±0.06	0.57±0.05	1.03±0.07
Maximum Flowering	MF	3	46.0	59.0	0.80±0.08	0.71±0.08	1.12±0.06
Final Flowering	FF	3	52.0	64.7	0.87±0.07	0.53±0.07	0.97±0.07
Duration of Flowering	DF	3	12.7	10.0	0.93±0.11	0.08±0.15	0.58±0.13
Maturation Date	MD	2	223.5	214.5	0.61±0.07	0.88±0.08	
Duration of Maturation	DM	2	170.5	150.0	0.56±0.08	0.89±0.12	
Percentage of Failures	NF	2	2.0	0.6	0.60±0.29	0.27±0.18	
In-shell Weight	SW	2	3.4	2.8	1.25±0.13	0.90±0.10	
Kernel Weight	KW	2	1.0	1.0	0.78±0.18	0.60±0.12	
Percentage of Kernels	PK	2	30.6	35.7	0.55±0.14	0.54±0.11	
Number of Double Kernels	ND	2	0.4	3.0	1.19±0.45	0.19±0.38	
Flower Density	FD	3	4.3	3.3	0.72±0.23	0.65±0.16	0.24±0.27
Density of Yield	DY	3	4.3	3.0	0.94±0.15	0.42±0.12	0.00±0.19
Auto compatibility	AC	1	NO	YES			
Rugosity of the seed	RS	2	1.0	2.5	0.56±0.11	0.64±0.11	
Color of the seed	CS	2	2.5	1.5	0.26±0.14	0.37±0.13	

Table 5.- Probabilities of the chi-square statistic calculated to compare the distributions of traits regarding the marker genotypes. *indicates probability lower than 0.05.

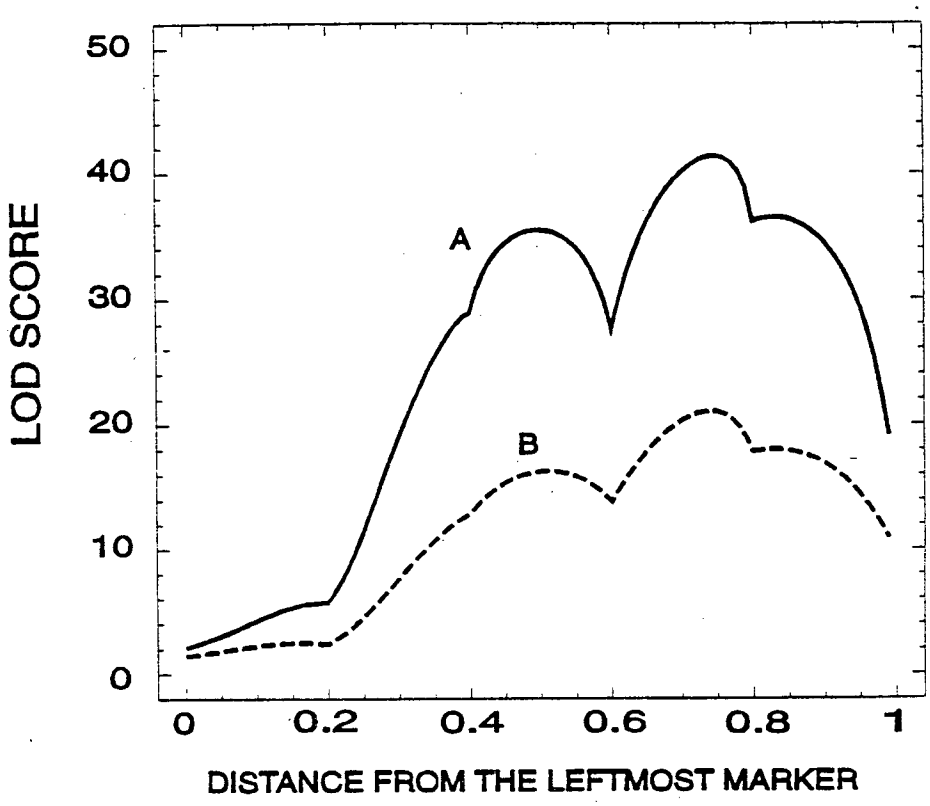
		Traits				
Marker	Year	FD	DY	AC	RS	CS
Pgm-2	1989	0.27	0.68	0.28	0.75	0.02*
	1990	0.47	0.49		0.64	0.11
	1991	0.28	0.93			
Pgi-2	1989	0.79	0.57	0.96	0.20	0.07
	1990	0.81	0.40		0.93	0.13
	1991	0.25	0.35			
Prx_c-2	1989	0.58	0.10	0.23	0.02*	0.13
	1990	0.15	0.03*		0.27	0.03
	1991	0.69	0.74			
Got-1	1989	0.34	0.97	0.89	0.46	0.36
	1990	0.91	0.68		0.14	0.15
	1991	0.62	0.30			
Got-2	1989	0.77	0.49	0.11	0.31	0.04*
	1990	0.80	0.21		0.18	0.23
	1991	0.54	0.73			

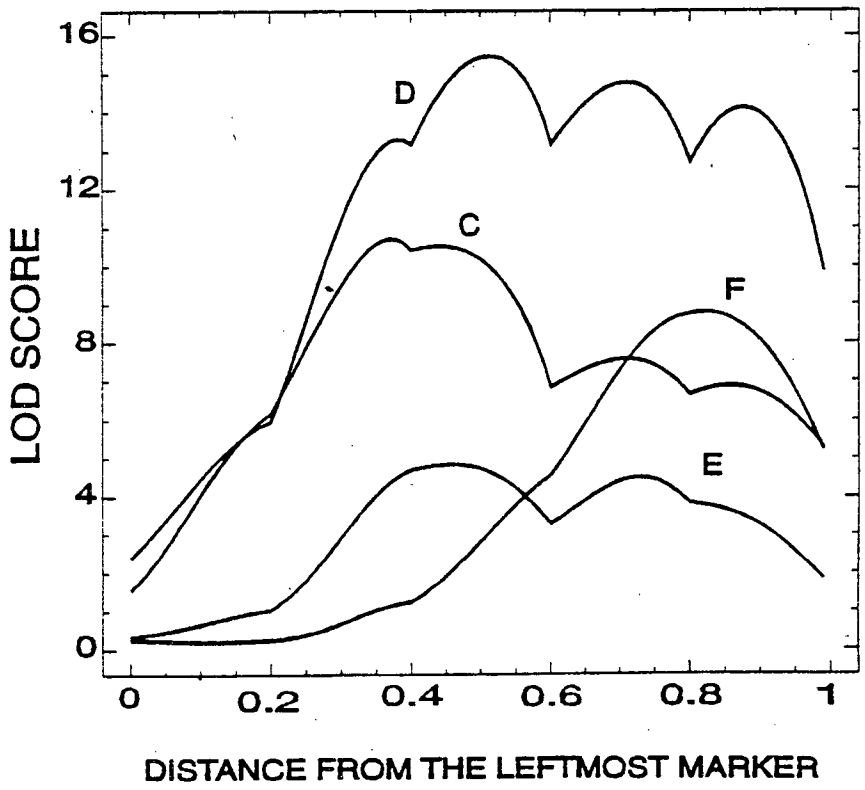
Table 6.- Probabilities of the F statistic obtained from the one-way ANOVA to test the association of the quantitative trait with a marker. * and ** indicate probability lower than 0.05 and 0.01, respectively.

		IF	MF	FF	DF	MD	DM	NF	SW	KW	PK	ND
Pg -2	1989	0.07	0.32	0.73	0.23	0.16	0.12	0.18	0.17	0.03*	0.89	0.23
	1990	0.05*	0.10	0.51	0.06	0.57	0.45	0.61	0.28	0.02*	0.68	0.12
	1991	0.58	0.11	0.25	0.60							
Pgi-2	1989	0.08	0.06	0.72	0.23	0.21	0.16	0.87	0.05*	0.04	0.39	0.73
	1990	0.27	0.20	0.48	0.52	0.85	0.92	0.82	0.03*	0.27	0.13	0.74
	1991	0.25	0.15	0.28	0.51							
Prx_c-2	1989	0.79	0.33	0.51	0.25	0.21	0.13	0.49	0.72	0.40	0.56	0.21
	1990	0.36	0.33	0.13	0.51	0.34	0.70	0.23	0.08	0.02*	0.01**	0.81
	1991	0.00**	0.01**	0.02*	0.04*							
Got-1	1989	0.21	0.34	0.07	0.36	0.89	0.51	0.58	0.64	0.20	0.81	0.18
	1990	0.18	0.08	0.00**	0.07	0.95	0.34	0.60	0.92	0.27	0.99	0.03*
	1991	0.05*	0.23	0.11	0.11							
Got-2	1989	0.04*	0.06	0.00**	0.08	0.48	0.06	0.58	0.41	0.86	0.25	0.30
	1990	0.00**	0.00**	0.00**	0.91	0.22	0.03*	0.39	0.70	0.90	0.35	0.88
	1991	0.51	0.84	0.96	0.22							

Figure Legends

Figure 1. Typical results found for two linked QTLs actually located at 42 and 82 cM from the leftmost marker. Curve F involving a single QTL at 82 cM is included for comparison purposes.


LOD SCORE
50
40
30
20
10
0
A
B
0
0.2
0.4
0.6
0.8
1
DISTANCE FROM THE LEFTMOST MARKER


LOD SCORE
16
12
8
4
0
D
C
F
E
0
0.2
0.4
0.6
0.8
1
DISTANCE FROM THE LEFTMOST MARKER

CATEGORICAL DATA ANALYSIS IN BIOTECHNOLOGY RESEARCH

E. Barry Moser and Raúl E. Macchiavelli
Dept. of Experimental Statistics, Louisiana State University

Paper presented at the meeting "Statistical Methods in Environmental Research and Molecular Biology", Cali, Colombia, June 1995

Abstract

Categorical responses are very common in biotechnology research. In this paper we present the main ideas of generalized linear models, with particular emphasis to logit models. We apply these models to two practical situations. The first problem studies the effect of chilling on bud development in azalea flowers, and the second situation analyzes the relationship between a genetic polymorphism in certain protein and obesity in humans. Both problems are analyzed with SAS PROC CATMOD.

Resumen

Respuestas medidas en una escala categórica aparecen comúnmente en la investigación biotecnológica. En este trabajo se presentan las ideas principales de los modelos lineales generalizados, enfatizando los modelos *logit*. Se presentan dos aplicaciones prácticas. En la primera se estudia el efecto del tratamiento de frío sobre el desarrollo floral en plantas de azalea. El segundo problema estudia la relación entre un polimorfismo genético en cierta proteína y obesidad en humanos. Ambos problemas se analizan en PROC CATMOD en SAS.

1. Introduction

The linear model for normally distributed random variables is very commonly used to address hypotheses concerning the mean or expected value of the random variables. Let y_{ij} be a normally distributed response measured on the jth individual of group i. Then y_{ij} could be modeled as

$$y_{ij} = \mu_i + \epsilon_{ij} \text{ for } i = 1,2,\cdots,I; j = 1,2,\cdots,n_i \tag{1}$$

where μ_i is the ith group mean, ϵ_{ij} is the error or amount the jth individual of group i differs from its mean, I is the number of groups, and n_i is the number of individuals in group i. Hypotheses of interest often include the null hypothesis H_0: $\mu_1 = \cdots = \mu_I$ and in general, these hypotheses can be tested using linear contrasts of the parameters. Alternative parameterizations of model (1) can be formulated and may lead to direct estimation of desired parameters and testing of desired hypotheses by the software employed. In fact a common parameterization of model (1) is the "effects" model

$$\mu_i = \mu + \alpha_i \text{ and } \sum_{i=1}^{I} \alpha_i = 0 \tag{2}$$

where μ serves as an "overall" mean and the α_i are the amounts by which the individual factor level means differ from the "overall" mean.

In categorical data analysis, we often observe the number of items out of a total that fall into each of several mutually exclusive categories K. Based upon random sampling principles these data are often treated as a sample from a multinomial distribution, although other distributions such as the Poisson and the negative binomial distribution are also used. It is tempting to model the category counts using the normal theory model proposed above. However, we must account for several qualities of the categorical data, and the normal theory approach may not adequately address them. First, we must define what the parameters of interest are. For multinomial data, the parameters of interest would be the proportion of individuals in the population that belong to each category or the probability that a

randomly selected individual would belong to a given category. We must also account for the fact that the proportions or probabilities of the mutually exclusive categories must sum to 1. Further, we would like to restrict any predictions or estimates of proportions or probabilities to the interval [0,1].

A general class of models called generalized linear models (see e.g., Dobson 1983, Agresti 1990) have been developed to address issues such as those raised above for multinomial models. In generalized linear models, the model parameters enter the equation in a linear, additive manner, such as proposed in (1) and (2) above, yet the distribution of the response is accounted for through the estimation method and through the way in which the mean is included in the model (the link function). Generalized linear models address distributions that belong to the exponential family of distributions. A distribution belongs to the exponential family if it is of the form

$$f(y;\theta) = \exp(a(y)b(\theta) + c(\theta) + d(y)). \tag{3}$$

For Y distributed as a normal(μ, σ^2) random variable we have

$$f(y;\theta) = \exp(y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}). \tag{4}$$

Notice that σ^2 is treated as a nuisance parameter and is assumed to be known. For Y distributed binomial(n, π),

$$f(y;\theta) = \exp(y\log\pi - y\log(1-\pi) + n\log(1-\pi) + \log\binom{n}{y}) \tag{5}$$

or

$$f(y;\theta) = \exp(y\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y}). \tag{6}$$

If $a(y) = 1$ as it does for both the normal and binomial distributions, then the distribution is said to be

in the canonical form and $b(\theta)$ is called the natural parameter of the distribution. In generalized linear models it is common to model the natural parameter as a linear function of the model predictors. Thus, the transformation specified by the natural parameter is applied to the observed means, which are then modeled as a linear function of the model parameters. For the normal distribution, the response function F is taken to be μ/σ^2, or usually just μ, and is modeled as a linear function of the model parameters. Remember that σ^2 is assumed known. In practice it is estimated from the error in fit. For the binomial distribution, the response function F is taken to be the logit transformation,

$$F = \log\left(\frac{\pi}{1-\pi}\right) \tag{7}$$

and is modeled as a linear function, say, of the groups,

$$F_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mu + \alpha_i \tag{8}$$

as we did before with the normal theory data, where π_i is the probability of being in the ith group. The parameters of these models are usually estimated using maximum likelihood although alternatives based upon weighted least squares and chi-square goodness-of-fit criteria are also available (see e.g., Agresti 1990:445-477). If the parameters are estimated using maximum likelihood, then likelihood ratio and Wald tests can be used to test hypotheses about the model parameters. Pearson chi-square and likelihood ratio goodness-of-fit tests can also be used to evaluate model adequacy for restricted models.

Assume that we have observed a binomial response from each of two populations and we would like to test the null hypothesis that H_0: $\pi_1 = \pi_2$. This hypothesis implies the hypothesis H_0: $F_1 = F_2$ (or H_0: $F_1 - F_2 = 0$). Thus, we can consider the functions F_1 and F_2 and construct the linear hypothesis

$$F_1 - F_2 = \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_2}{1-\pi_2}\right) = \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right) \tag{9}$$

which is simply the log-odds ratio, the commonly reported statistic in contingency table analysis (Fienberg 1980:17).

If there are more than 2 groups, and these groups, for example, are arranged in a factorial treatment arrangement, then additional model parameters can be specified to focus upon the main effects and interactions of the treatment, experimental, and/or observational factors. For a 2-factor factorial treatment arrangement, the response function could be modeled as

$$F_{ij} = \log\left(\frac{1-\pi_{ij}}{\pi_{ij}}\right) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{10}$$

with $\sum_{i=1}^{I} \alpha_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$, $\sum_{i=1}^{I} (\alpha\beta)_{ij} = 0 \quad \forall j$, $\sum_{j=1}^{J} (\alpha\beta)_{ij} = 0 \quad \forall i$ where the α_i compose the main effect of factor 1 for $i = 1,2,\ldots,I$ levels, the β_j compose the main effect of factor 2 for $j = 1,2,\ldots,J$ levels, and the $(\alpha\beta)_{ij}$ compose the interaction of factor 1 with factor 2. The hypothesis H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$ tests for homogeneity of proportions across the levels of factor 1. The hypothesis H_0: $\beta_1 = \beta_2 = \ldots = \beta_J = 0$ tests for homogeneity of proportions across the levels of factor 2. While the hypothesis H_0: $(\alpha\beta)_{ij} = 0 \quad \forall \quad (i,j),$, tests that the distribution of proportions across the levels of factor 1 is the same across each of the levels of factor 2.

If more than 2 response categories exist, then the logit transformation can be generalized in several ways. In this paper we will focus on categories that are ordered. For an ordinal response the cumulative logit transformation is commonly applied. Let i refer to one of the I response categories, then this transformation computes the log of the odds of being in a category larger than i to that of being in or below category i. If we have sampled from J groups, then the response functions become

$$F_{ij} = \log\left(\frac{1-\sum_{k=1}^{i}\pi_k}{\sum_{k=1}^{i}\pi_k}\right). \qquad (11)$$

Since there are now more than 2 categories, we will need to add a modeling term, z_i, corresponding to the level of categories being compared in the odds ratio. This permits the odds to depend upon which categories are being compared. Now our model for the response functions can take the form

$$F_{ij} = \mu + z_i + \beta_j \qquad (12)$$

where z_i behaves much like a covariate in an analysis of covariance model.

The categorical models developed in (10) and (12) will now be illustrated through 2 biotechnology experiments.

2. Chilling and Bud Development in Azaleas

A continuing problem in azalea production for flowering potted plants has been uniform flowering. During normal production cycle, azaleas will not bloom unless flower bud dormancy is overcome. Traditionally, breaking dormancy of azalea flower buds is accomplished by exposing plants to temperatures of 3-10°C (chilling) through natural cooling or refrigerated storage.

In this data set, obtained by Rebecca Moss (Dept. of Horticulture, Louisiana State University), the developmental stage of flower buds from plants which were chilled and non-chilled were

recorded for three different shipments. The stages were classified in two categories (b,c). (Other shipments also included category a.)

The observed frequencies are presented in table 1. There are 12 cells.

Table 1: Observed frequencies of buds in two developmental stages. Data provided by R. Moss.

Treatment	Shipment	Stage b	Stage c
non-chilled	1	166	150
non-chilled	2	231	102
non-chilled	3	370	7
chilled	1	202	196
chilled	2	145	147
chilled	3	360	3

Let π_{ij} be the probability that a bud will be in stage b (in a plant from shipment j treated with treatment i). Since the response has only two levels (stage b or c), the response function to consider is

$$F_{ij} = \log \frac{\pi_{ij}}{1-\pi_{ij}},$$

for i=1,2 (treatments) and j=1,2,3 (shipments). F_{ij} is the logarithm of the odds in favor of a bud being in stage b (instead of stage c) in a plant from treatment i and shipment j.

The saturated model will be a model with one parameter for each F_{ij} (or equivalently, for each π_{ij}). The fit for this model is trivially exact (no goodness of fit test is possible).

Notice that any model with 6 or more linearly independent parameters will be equivalent to this one. For example,

$$F_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}, \tag{1}$$

with the usual restrictions $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \alpha\beta_{ij} = \sum_j \alpha\beta_{ij} = 0$ has six parameters and is equivalent to the saturated one.

In order to fit this model using SAS PROC CATMOD, the following program can be used:

proc catmod;
model nwstg= trt shp trt*shp / freq pred prob;

The output obtained will include the frequency table presented as Table 1 (requested by the option **freq**), the observed probabilities (or relative frequency table, requested by **prob**), information about the iterative estimation procedure, an "analysis of variance table" presenting chi-squared tests for the main hypotheses of interest, and a list of estimated parameters. See Appendix A for a complete output.

A simpler model (with main effects only) was tried with the commands:

proc catmod;
model nwstg= trt shp / freq pred prob;

This model has only 4 parameters, leaving 2 degrees of freedom to test the goodness of fit. The likelihood ratio statistic for this test was 14.91, with a p-value of .0006. Therefore this model does not fit well, and the interaction term should be included (i.e., the saturated model will be used).

Like in ANOVA models, the presence of interactions makes the tests about main effects not very interesting. Contrasts can be used to test the hypotheses of interest (like simple effects). In this problem, 4 hypotheses were identified as important to the researcher. They were:

- Does chilling have an effect in shipment 1?
- Does chilling have an effect in shipment 2?
- Does chilling have an effect in shipment 3?
- Do non-chilled plants differ among shipments?

The last one will be a contrast with 2 df, while all the others will have 1 df each. In order to write the CONTRAST statements in SAS, we need to write them in terms of CATMOD full rank parametrization (which has "sum-to-zero" restrictions for all parameters).

For example, the first contrast is

$$H_0: F_{11} - F_{21} = 0.$$

Writing this contrast in terms of the parameters in the model (1),

$$\mu + \alpha_1 + \beta_1 + \alpha\beta_{11} - (\mu + \alpha_2 + \beta_1 + \alpha\beta_{21}) = 0.$$

Since $\alpha_2 = -\alpha_1$ and $\alpha\beta_{21} = -\alpha\beta_{11}$, the contrast is

$$2\alpha_1 + 2\alpha\beta_{11} = 0.$$

The other two simple effects of treatment can be obtained similarly. Notice that this form is not a very familiar one for a contrast, but it can be written directly in SAS:

```
contrast 'trt in shp1' trt 2 trt*shp 2 0;
contrast 'trt in shp2' trt 2 trt*shp 0 2;
contrast 'trt in shp3' trt 2 trt*shp -2 -2;
```

The last contrast (shipment effect in non-chilled plants) can be written as

$$H_0: F_{11} - F_{13} = 0 \text{ and } F_{12} - F_{13} = 0.$$

The corresponding SAS statement is

```
contrast 'nonchilled' shp 2 -1 trt*shp 2 -1, shp -1 2 trt*shp 1 -2;
```

The output for these contrasts are likelihood ratio tests:

CONTRASTS OF MAXIMUM-LIKELIHOOD ESTIMATES

Contrast	DF	Chi-Square	Prob
trt in shp1	1	0.22	0.6368
trt in shp2	1	24.82	0.0000
trt in shp3	1	1.40	0.2373
nonchilled	2	95.27	0.0000

From these results we can interpret the interaction very clearly: while chilling does not seem to have

an effect in shipments 1 and 3, it does have a significant effect in shipment 2. Also, there are differences between shipments in the non treated plants.

3. Effects of genetic polymorphism in APO-AIV

In this example we will use CATMOD to model ordinal responses. The study involves determining the effect of genetic polymorphism in apolipoprotein AIV (APO-AIV) on body mass index in human subjects. The body mass index is defined as

$$BMI = \frac{\text{weight } (kg)}{\text{height }^2 (\text{in } \text{m}^2)}.$$

APO-AIV is a protein associated with HDL (high density lipoprotein) in plasma and as "free" APO-AIV. It might have a role in reverse cholesterol metabolism or in triglycerides.

Several phenotypes have been identified for APO-AIV. Two of them were the most frequent and were used in this analysis (they were coded as 0, 1). The data were obtained by Dr. Michael Lefevre (Pennington Biomedical Research Center, Louisiana State University).

The problem is to check how these phenotypes differ on BMI. Since this variable, representing how "obese" or "lean" a person is, also depends on several other factors (like age, sex, race, etc.), the researcher did not use the actual BMI, but the BMI percentile. This BMI percentile was obtained from the raw BMI score by looking at tables constructed for each sex, race and age group. This is a way of eliminating the effects of these factors, transforming the response into an ordinal variable. The percentiles considered were 5, 10, 15, 20, 50, 75, 85, 90, 95 and 99.

The research question in this study is "do people in phenotype 1 tend to be in a higher BMI percentile than people in phenotype 0?". This question can be addressed by modelling the *cumulative logits*

(CLOGIT option). Let π_{ij} be the probability that a subject will be in the *i*-th ordered class and the *j*-th group. The cumulative logit F_{ij} is defined as the logarithm of the odds in favor of being in a class larger than *i*:

$$F_{ij} = \log \frac{1 - \sum_{k=1}^{i} \pi_{kj}}{\sum_{k=1}^{i} \pi_{kj}} \qquad (2)$$

The linear model for this F_{ij}'s is

$$F_{ij} = \mu + \tau_i + \beta_j,$$

where τ represents the percentile effect and β the phenotype effect. The usual restrictions apply ($\sum \tau_i = \sum \beta_j = 0$). This model can be fitted with the following SAS program:

```
proc catmod data=colo.bmi;
response clogits;
model bmi_per = _response_ pheno / freq predict prob;
weight tot;
```

The results of this analysis are presented as Appendix B. The first thing to notice is the goodness of fit test, which has 8 d.f. In the output it is found under "residual". The χ^2 statistic is 4.75, with a p-value of .7835. From this test, one concludes that the fit is good, and hence we can interpret the parameter estimates. Another part of the output to check for goodness of fit is the listing of residuals (obtained with the **predict** option). There is certain pattern of negative residuals for low function

numbers in PHENO=0 and high function numbers in PHENO=1. This is an effect of requiring a

constant odds-ratio in the model (which we will discuss later).

The main research question (PHENO effect) can be answered by looking at the χ^2 statistic for PHENO. Its value is 5.48, with 1 d.f. and a p-value of .0192. Therefore, there is a significant effect of phenotype on BMI. In order to interpret the nature of this effect, we need to look at $\hat{\beta} = .23$. (Notice that there is only one parameter associated with PHENO, since $\beta_2 = -\beta_1$.)

The difference of the functions associated with the same percentile at the two values of PHENO is

$$F_{i1} - F_{i2} = \mu + \tau_i + \beta - (\mu + \tau_i - \beta) = 2\beta.$$

If we write the definition of the cumulative logits we can conclude that

$$\log \frac{\frac{1-\sum_{k=1}^{i} \pi_{k1}}{\sum_{k=1}^{i} \pi_{k1}}}{\frac{1-\sum_{k=1}^{i} \pi_{k2}}{\sum_{k=1}^{i} \pi_{k2}}} = 2\beta.$$

This equation is saying that the odds ratio (the ratio of the odds in favor of being in a class larger than *i* for PHENO=0 and 1) is a constant:

$$\frac{\text{odds for PHENO = 0}}{\text{odds for PHENO = 1}} = \exp(2\beta).$$

Notice that this does not depend on the particular percentile (i). The estimated odds-ratio is obtained from $\hat{\beta}$:

$$\exp(2\hat{\beta}) = 1.58$$

From this relationship, we estimate that the odds in favor of being in a class larger than i (for any i) are 1.58 times larger in PHENO=0 than in PHENO=1. Hence, people with PHENO=0 are more likely to be in the higher percentiles (i.e., being more "obese") than people with PHENO=1.

In order to obtain the fitted probabilities we can proceed as follows. Consider PHENO=0.

$$\hat{F}_{11} = \log\frac{1-\hat{\pi}_{11}}{\hat{\pi}_{11}} = \hat{\mu} + \hat{\tau}_1 + \hat{\beta} = 1.0382 + 2.9725 + .2300 = 4.2407 .$$

Solving this equation for $\hat{\pi}_{11}$ we obtain

$$\hat{\pi}_{11} = \frac{1}{1 + e^{4.2407}} = .0142 .$$

Similarly,

$$\hat{F}_{11} = \log\frac{1-\hat{\pi}_{11}-\hat{\pi}_{21}}{\hat{\pi}_{11}+\hat{\pi}_{21}} = \hat{\mu} + \hat{\tau}_2 + \hat{\beta} = 1.0382 + 2.1580 + .2300 = 3.4262 .$$

From this equation,

$$\hat{\pi}_{11} + \hat{\pi}_{21} = \frac{1}{1 + e^{3.4262}} = .0315,$$

and $\hat{\pi}_{21} = .0173$. Continuing in this fashion we can obtain all fitted probabilities.

References

Agresti, A. (1990) *Categorical Data Analysis*. John Wiley & Sons, New York.

Dobson, A. J. (1983). *An Introduction to Statistical Modelling*. Chapman and Hall, London.

Fienberg, S. E. (1980). *The Analysis of Cross-classified Categorical Data*. Second edition. The MIT Press, Cambridge, MA.

SAS Institute, Inc. (1990). *SAS/STAT User's Guide, Version 6*. Fourth edition, Volume 1. SAS Institute, Inc., Cary, NC.

A. Azalea Example: SAS Output

CATMOD PROCEDURE

Response: NWSTG
Weight Variable: None
Data Set: A
Frequency Missing: 0

Response Levels (R)= 2
Populations (S)= 6
Total Frequency (N)= 2079
Observations (Obs)= 2079

POPULATION PROFILES

Sample	TRT	SHP	Sample Size
1	1	4	316
2	1	5	333
3	1	6	377
4	3	4	398
5	3	5	292
6	3	6	363

RESPONSE PROFILES

Response	NWSTG
1	b
2	c

RESPONSE FREQUENCIES

Sample	Response Number 1	2
1	166	150
2	231	102
3	370	7
4	202	196
5	145	147
6	360	3

RESPONSE PROBABILITIES

Sample	Response Number 1	2
1	0.52532	0.47468
2	0.69369	0.30631
3	0.98143	0.01857
4	0.50754	0.49246
5	0.49658	0.50342
6	0.99174	0.00826

MAXIMUM-LIKELIHOOD ANALYSIS

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion
0	0	2882.106	1.0000
1	0	2056.5688	0.2864
2	0	1951.538	0.0511
3	0	1927.222	0.0125
4	0	1923.5425	0.001909
5	0	1923.3692	0.0000901
6	0	1923.3686	3.1756E-7
7	0	1923.3686	4.598E-12

Non Interaction Model

	Parameter Estimates			
Iteration	1	2	3	4
0	0	0	0	0
1	0.8059	0.1235	-0.7301	-0.4076
2	1.1519	0.1761	-1.0695	-0.7485
3	1.4079	0.1894	-1.3238	-1.0046
4	1.5540	0.1908	-1.4697	-1.1508
5	1.5947	0.1909	-1.5105	-1.1915
6	1.5973	0.1909	-1.5131	-1.1941
7	1.5973	0.1909	-1.5131	-1.1941

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	201.41	0.0000
TRT	1	12.10	0.0005
SHP	2	167.10	0.0000
LIKELIHOOD RATIO	2	14.91	0.0006

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	1.5973	0.1126	201.41	0.0000
TRT	2	0.1909	0.0549	12.10	0.0005
SHP	3	-1.5131	0.1206	157.51	0.0000
	4	-1.1941	0.1222	95.51	0.0000

MAXIMUM-LIKELIHOOD PREDICTED VALUES FOR RESPONSE FUNCTIONS AND PROBABILITIES

		--------Observed-------		---------Predicted-------		
Sample	Function Number	Function	Standard Error	Function	Standard Error	Residual
1	1	0.10135249	0.11265329	0.27517417	0.09723315	-0.1738217
	P1	0.52531646	0.02809112	0.56836271	0.02385387	-0.0430463
	P2	0.47468354	0.02809112	0.43163729	0.02385387	0.04304626
2	1	0.8174449	0.11888198	0.59415953	0.09787989	0.22328537
	P1	0.69369369	0.02526037	0.64431896	0.02243133	0.04937473
	P2	0.30630631	0.02526037	0.35568104	0.02243133	-0.0493747
3	1	3.96759286	0.38152306	4.49543204	0.32481564	-0.5278392
	P1	0.98143236	0.00695245	0.98896331	0.00354532	-0.0075309
	P2	0.01856764	0.00695245	0.01103669	0.00354532	0.00753095
4	1	0.03015304	0.10026234	-0.10666	0.08930642	0.13681301
	P1	0.50753769	0.02505989	0.47336026	0.02226323	0.03417743
	P2	0.49246231	0.02505989	0.52663974	0.02226323	-0.0341774
5	1	-0.0136988	0.11704389	0.21232539	0.09957698	-0.2260242
	P1	0.49657534	0.0292596	0.55288282	0.02461577	-0.0563075
	P2	0.50342466	0.0292596	0.44711718	0.02461577	0.05630748
6	1	4.78749174	0.5797509	4.1135979	0.32173498	0.67389384
	P1	0.99173554	0.00475173	0.98391414	0.00509213	0.0078214
	P2	0.00826446	0.00475173	0.01608586	0.00509213	-0.0078214

Saturated Model

MAXIMUM-LIKELIHOOD ANALYSIS

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion
0	0	2882.106	1.0000
1	0	2040.5745	0.2920
2	0	1937.3417	0.0506
3	0	1912.897	0.0126
4	0	1908.7759	0.002154
5	0	1908.4615	0.000165
6	0	1908.4568	2.4755E-6
7	0	1908.4567	8.935E-10

Parameter Estimates

Iteration	1	2	3	4	5	6
0	0	0	0	0	0	0
1	0.7975	0.1364	-0.7318	-0.4170	-0.1008	0.2578
2	1.1449	0.1294	-1.0791	-0.7432	-0.0938	0.2860
3	1.4013	0.0990	-1.3355	-0.9994	-0.0634	0.3166
4	1.5550	0.0529	-1.4892	-1.1531	-0.0173	0.3627
5	1.6082	0.0200	-1.5424	-1.2063	0.0156	0.3956
6	1.6149	0.0139	-1.5492	-1.2131	0.0217	0.4017
7	1.6151	0.0137	-1.5493	-1.2132	0.0219	0.4018

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	176.51	0.0000
TRT	1	0.01	0.9100
SHP	2	149.12	0.0000
TRT*SHP	2	14.50	0.0007
LIKELIHOOD RATIO	0	.	.

Saturated Model

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	1.6151	0.1216	176.51	0.0000
TRT	2	0.0137	0.1216	0.01	0.9100
SHP	3	-1.5493	0.1291	143.97	0.0000
	4	-1.2132	0.1308	86.09	0.0000
TRT*SHP	5	0.0219	0.1291	0.03	0.8656
	6	0.4018	0.1308	9.44	0.0021

MAXIMUM-LIKELIHOOD PREDICTED VALUES FOR RESPONSE FUNCTIONS AND PROBABILITIES

		--------Observed---------		--------Predicted---------		
Sample	Function Number	Function	Standard Error	Function	Standard Error	Residual
1	1	0.10135249	0.11265329	0.10135249	0.11265329	0
	P1	0.52531646	0.02809112	0.52531646	0.02809112	0
	P2	0.47468354	0.02809112	0.47468354	0.02809112	0
2	1	0.8174449	0.11888198	0.8174449	0.11888198	0
	P1	0.69369369	0.02526037	0.69369369	0.02526037	0
	P2	0.30630631	0.02526037	0.30630631	0.02526037	0
3	1	3.96759286	0.38152306	3.96759286	0.38152275	0
	P1	0.98143236	0.00695245	0.98143236	0.00695244	0
	P2	0.01856764	0.00695245	0.01856764	0.00695244	0
4	1	0.03015304	0.10026234	0.03015304	0.10026234	0
	P1	0.50753769	0.02505989	0.50753769	0.02505989	0
	P2	0.49246231	0.02505989	0.49246231	0.02505989	0
5	1	-0.0136988	0.11704389	-0.0136988	0.11704389	0
	P1	0.49657534	0.0292596	0.49657534	0.0292596	0
	P2	0.50342466	0.0292596	0.50342466	0.0292596	0
6	1	4.78749174	0.5797509	4.78749146	0.57953514	2.81698E-7
	P1	0.99173554	0.00475173	0.99173553	0.00474996	2.30884E-9
	P2	0.00826446	0.00475173	0.00826447	0.00474996	-2.3088E-9

CONTRASTS OF MAXIMUM-LIKELIHOOD ESTIMATES

Contrast	DF	Chi-Square	Prob
nonchilled	2	95.27	0.0000
trt in shp4	1	0.22	0.6368
trt in shp5	1	24.82	0.0000
trt in shp6	1	1.40	0.2373

B. APO-AIV Example: SAS Output

CATMOD PROCEDURE

Response: BMI_PER	Response Levels (R)= 10
Weight Variable: TOT	Populations (S)= 2
Data Set: BMI	Total Frequency (N)= 615
Frequency Missing: 0	Observations (Obs)= 20

POPULATION PROFILES

Sample	PHENO	Sample Size
1	0	520
2	1	95

RESPONSE PROFILES

Response	BMI_PER
1	5
2	10
3	15
4	25
5	50
6	75
7	85
8	90
9	95
10	99

RESPONSE FREQUENCIES

	Response Number				
Sample	1	2	3	4	5
1	8	10	6	28	76
2	1	1	2	8	21

	Response Number				
Sample	6	7	8	9	10
1	105	74	55	77	81
2	18	16	12	8	8

RESPONSE PROBABILITIES

Sample	Response Number 1	2	3	4	5
1	0.01538	0.01923	0.01154	0.05385	0.14615
2	0.01053	0.01053	0.02105	0.08421	0.22105

Sample	Response Number 6	7	8	9	10
1	0.20192	0.14231	0.10577	0.14808	0.15577
2	0.18947	0.16842	0.12632	0.08421	0.08421

Sample	Function Number	Response Function
1	1	4.15888
	2	3.32823
	3	3.02852
	4	2.19722
	5	1.11923
	6	0.20844
	7	-0.36556
	8	-0.82905
	9	-1.69005
2	1	4.54329
	2	3.83945
	3	3.12457
	4	1.93393
	5	0.63063
	6	-0.14764
	7	-0.87249
	8	-1.59686
	9	-2.38647

Sample	Function Number	DESIGN MATRIX 1	2	3	4	5	6	7	8	9	10
1	1	1	1	0	0	0	0	0	0	0	1
	2	1	0	1	0	0	0	0	0	0	1
	3	1	0	0	1	0	0	0	0	0	1
	4	1	0	0	0	1	0	0	0	0	1
	5	1	0	0	0	0	1	0	0	0	1
	6	1	0	0	0	0	0	1	0	0	1
	7	1	0	0	0	0	0	0	1	0	1
	8	1	0	0	0	0	0	0	0	1	1
	9	1	-1	-1	-1	-1	-1	-1	-1	-1	1
2	1	1	1	0	0	0	0	0	0	0	-1
	2	1	0	1	0	0	0	0	0	0	-1
	3	1	0	0	1	0	0	0	0	0	-1
	4	1	0	0	0	1	0	0	0	0	-1
	5	1	0	0	0	0	1	0	0	0	-1
	6	1	0	0	0	0	0	1	0	0	-1
	7	1	0	0	0	0	0	0	1	0	-1
	8	1	0	0	0	0	0	0	0	1	-1
	9	1	-1	-1	-1	-1	-1	-1	-1	-1	-1

ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	1	73.89	0.0000
RESPONSE	8	718.04	0.0000
PHENO	1	5.48	0.0192
RESIDUAL	8	4.75	0.7835

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	1.0382	0.1208	73.89	0.0000
RESPONSE	2	2.9725	0.2656	125.26	0.0000
	3	2.1580	0.1511	204.02	0.0000
	4	1.8361	0.1243	218.06	0.0000
	5	0.9614	0.0945	103.58	0.0000
	6	-0.1554	0.0885	3.08	0.0792
	7	-1.0428	0.0902	133.61	0.0000
	8	-1.6358	0.0937	304.98	0.0000
	9	-2.1206	0.0998	451.16	0.0000
PHENO	10	0.2300	0.0982	5.48	0.0192

PREDICTED VALUES FOR RESPONSE FUNCTIONS

Sample	Function Number	-------Observed-------- Function	Standard Error	--------Predicted-------- Function	Standard Error	Residual
1	1	4.15888308	0.35630482	4.24069836	0.33545855	-0.0818153
	2	3.32822836	0.23989078	3.42617757	0.22710874	-0.0979492
	3	3.0285221	0.2090043	3.10430852	0.19600199	-0.0757864
	4	2.19722458	0.14617634	2.22958043	0.13692893	-0.0323559
	5	1.11923158	0.10180138	1.11280246	0.09796929	0.00642912
	6	0.20844376	0.08818257	0.22540855	0.08671575	-0.0169648
	7	-0.3655556	0.08917491	-0.3675762	0.08765506	0.00202059
	8	-0.8290492	0.09534959	-0.8523541	0.09351136	0.02330488
	9	-1.6900503	0.12092801	-1.7050802	0.11720611	0.0150299
2	1	4.54329478	1.00530508	3.78060518	0.37031224	0.76268961
	2	3.83945231	0.71466964	2.96608438	0.27571165	0.87336793
	3	3.12456515	0.51087084	2.64421533	0.25053362	0.48034981
	4	1.93393396	0.308839	1.76948724	0.20748665	0.16444671
	5	0.63062682	0.215481	0.65270927	0.18662758	-0.0220824
	6	-0.147636	0.20575499	-0.2346846	0.18434808	0.08704863
	7	-0.8724881	0.22503257	-0.8276694	0.18688966	-0.0448188
	8	-1.5968591	0.27415001	-1.3124472	0.19183472	-0.2844119
	9	-2.3864666	0.36945129	-2.1651733	0.20695884	-0.2212932