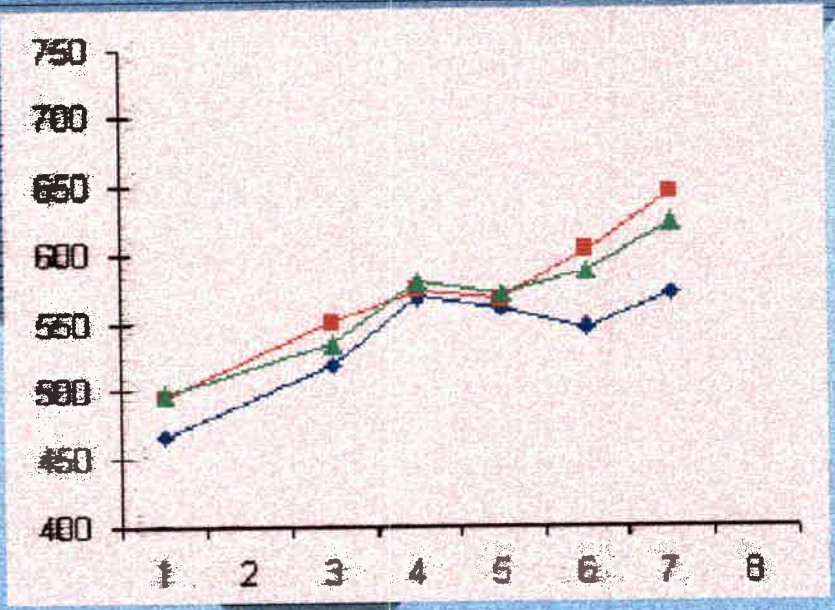
ESTRATEGIAS PARA EL ANÁLISIS DE ENSAYOS CON MEDIDAS REPETIDAS



Trabajo de Monografía correspondiente a la carrera de posgrado
Especialidad en Estadística - Orientación Diseño Experimental

Facultad de Ciencias Exactas, Físico - Químicas y Naturales

Universidad Nacional de Río Cuarto

1999

ESTRATEGIAS PARA EL ANÁLISIS DE ENSAYOS CON MEDIDAS REPETIDAS

Grado a obtener:	Especialista en Estadística Orientación en Diseño Experimental
Nombre del tesista:	Prof. María Gabriela Palacio
Directora:	Mg. María Inés Rodríguez
Jurado:	Mg. Héctor Agnelli Dra. Patricia Giménez Mg. Elsa Moschetti
Fecha:	Noviembre de 1999

A mi papá Mario que siempre está
incondicionalmente a mi lado,
a mi mamá Rosa que desde
el Cielo me acompaña y
a mi novio Gerardo
por su paciencia.

INDICE

Resumen

Las medidas repetidas son utilizadas en muchos campos de investigación, y son aún más comunes que las mediciones simples. El término "repetidas" hace referencia a la medición de la misma característica a una misma unidad experimental en más de una ocasión.

Este tipo de estudios, frecuentemente llamados longitudinales, presentan una estructura similar a la del análisis multivariado, pero combinada con elementos de series de tiempo. Se diferencian del clásico análisis de datos multivariados, porque su aspecto de serie de tiempo les imparte una estructura de interdependencia a los datos que no se presenta en los datos multivariados, y ellos difieren también de las clásicas series de tiempo pues consisten en un gran número de series cortas, una para cada individuo, más que en una sola serie larga de tiempo.

Para probar los efectos que involucran los factores con medidas repetidas, se realiza el análisis de la varianza entre-grupos usual. Si el factor tiene más de dos niveles puede probarse su significación con un análisis univariado o con uno multivariado. Ambos enfoques son abordados detallando las diferencias entre ellos en cuanto a sus supuestos, la sensibilidad de sus pruebas y las conclusiones obtenidas.

La presente monografía realiza una presentación de los métodos de análisis de la varianza (univariado y multivariado) y de regresión aplicados a datos provenientes de ensayos con medidas repetidas, cuando la distribución del error es normal. Como el interés está centrado en presentar un texto accesible para consulta de investigadores usuarios de la estadística, además del contexto teórico adecuado, se plantean problemas resueltos con algunos paquetes estadísticos de uso masivo, con su correspondiente interpretación.

Objetivos

Este trabajo pretende:

1. ayudar a los investigadores ("usuarios" potenciales de las medidas repetidas) a reconocer situaciones experimentales en las que se utilizan medidas repetidas, presentando algunos casos y mostrando la forma correcta de analizar estas experiencias;

2. presentar una recopilación de las distintas alternativas para analizar datos de medidas repetidas, asumiendo normalidad de los errores, para el enfoque de Análisis de la Varianza y para la Regresión;

3. brindar un contexto teórico sobre los supuestos y adecuación de los distintos tipo de abordaje;

4. presentar ejemplos prácticos resueltos usando distintos softwares y las conclusiones de los mismos.

CAPITULO 1
Introducción

En el contexto en el cual se trabajará, el término "repetidas" hace referencia a la medición de la misma característica (que se considera como variable respuesta) a una misma unidad experimental (sujeto, parcela, animal, árbol, etc.) en más de una ocasión (temporal o espacialmente hablando). Este tipo de observaciones o mediciones repetidas sobre una misma unidad, no se originan necesariamente porque se efectúen en tiempos distintos. Por ejemplo, una respuesta biomédica puede ser observada en T distintos lugares del cuerpo de un paciente, por lo que cada individuo brinda un vector de medidas respuesta, pero ellas representan la misma medición física, efectuada en distintas secuencias de tiempo o de lugar.

A menudo se asocia a las "medidas repetidas" con el estudio de curvas de crecimiento en los estudios biológicos o con los diseños de parcelas divididas. Sin embargo, éstas también son muy comunes en la mayoría de los campos científicos donde se usa la estadística (a veces aún más comunes que las mediciones simples). En estudios que se llevan a cabo a lo largo de cierto período, las condiciones generales en las que se desarrolla el experimento pueden variar, y por lo tanto suele ser más útil y eficiente registrar datos de los mismos sujetos en diferentes momentos que realizar una sola observación de cada unidad experimental.

Uno de los principales propósitos al usar medidas repetidas es controlar la diferencia entre los sujetos. En estos experimentos los efectos de los tratamientos para el sujeto i son medidos en relación a la respuesta media de ese sujeto para cada uno de los tratamientos y así la variabilidad debida a la diferencia entre los sujetos es eliminada del error experimental. En las ciencias del comportamiento donde los elementos que forman las poblaciones en estudio son generalmente personas, dada la diferencia en experiencia y conocimientos previos de las mismas, las respuestas a los tratamientos

pueden ser muy variables. Esta variabilidad a veces se debe más a diferencias entre las personas antes de realizar el experimento que al tratamiento en sí. La idea cuando se realizan ensayos con medidas repetidas es separar esta fuente de variación de los efectos de los tratamientos y del error experimental para aumentar la sensibilidad del experimento.

En la mayoría de los estudios donde se realizan mediciones simples se pueden hacer también medidas repetidas, salvo cuando la realización del experimento es de cierta manera destructiva. Si, en cambio, el experimento sólo altera el estado de la unidad experimental, puede conducirse un ensayo con medidas repetidas. Para esto, se supone que no hay "efecto residual" de un tratamiento en un período, sobre la respuesta a otro tratamiento en el período siguiente (en estudios que se realizan con drogas, debe transcurrir suficiente tiempo entre la administración de una y otra droga para que la misma sea "limpiada" del organismo). También deben tenerse en cuenta los posibles efectos del aprendizaje, pues las respuestas de un sujeto podrían mejorar, sólo porque se realizó antes la misma evaluación.

Las medidas repetidas se utilizan por diversas causas:

- las observaciones repetidas pueden ser el único medio para obtener las mediciones requeridas (por ejemplo el conteo de las ocurrencias de algún fenómeno);

- el interés puede estar centrado en la evolución de la respuesta dadas ciertas condiciones iniciales que pueden estar o no fijadas experimentalmente (las curvas de crecimiento constituyen uno de estos casos);

- el investigador puede querer comparar los efectos de la administración continua de algún tratamiento a través del tiempo;

- diferentes tratamientos pueden querer compararse en situaciones en las que la variabilidad entre unidades es un factor importante no controlado (para incrementar la precisión se requieren comparaciones intra-unidades de los tratamientos);

- se puede requerir estudiar los efectos totales de diferentes secuencias de tratamientos (como en estudios de rotación de cultivos en agricultura).

En estas experiencias nos enfrentamos a tres tipos de relaciones entre las unidades experimentales:

a) Diferentes unidades se asumen como una muestra, por lo tanto las respuestas de distintas unidades experimentales son independientes entre sí.

b) Las respuestas de una misma unidad tienden a estar más relacionadas que las provenientes de unidades diferentes. Parte de esta variabilidad puede controlarse usando covariables. Lo restante debe tenerse en cuenta como dependencia estocástica (tratando de definir una matriz de correlación para cada sujeto), que puede modelarse con algún tipo de análisis multivariado.

c) Cuando una variable continua, como tiempo o espacio, sirve para distinguir entre observaciones, aquellas obtenidas "con mayor cercanía" en la misma unidad experimental están usualmente más relacionadas; este es un segundo tipo de dependencia estocástica.

En resumen, lo que distingue primordialmente estos ensayos de otros estudios (tal vez más usuales en estadística) es que:

* la misma característica es medida a la misma unidad experimental más de una vez, lo cual implica que las respuestas no son independientes como en el análisis de regresión usual;

* la relación que en general existe entre observaciones de la *misma* variable medida a la misma unidad experimental distingue a las medidas repetidas de la mayoría de los métodos multivariados, pues en estos se trabaja con interdependencia entre *diferentes* variables respuesta;

* hay más de una unidad experimental involucrada, y así las respuestas no forman una simple serie de tiempo.

Algunos autores consideran los ensayos con medidas repetidas como casos particulares del Diseño en Bloques Completamente Aleatorizados: cada sujeto constituye un bloque y es observado bajo todos los tratamientos (ocasiones). Como cada sujeto se parece a sí mismo más que a cualquier otro, la máxima homogeneidad dentro de un bloque se da cuando cada uno actúa como su propio control, formando un bloque.

A continuación se muestran algunas situaciones problemáticas donde se usan medidas repetidas:

1.1 - Se quiere estudiar si existe diferencia en el crecimiento de ratas sometidas a tres tratamientos (un control y dos con agregados químicos al agua que toman). Se seleccionan 3 grupos de ratas, aplicando a cada grupo un tratamiento y registrando el peso corporal inicial de cada rata y también el peso corporal en las semanas 1, 2, 3 y 4 (contadas desde el inicio del tratamiento).

1.2 - Cuatro grupos de sujetos participan en un experimento médico. Grupo 1: control; Grupo 2: pacientes diabéticos sin complicaciones; Grupo 3: pacientes diabéticos con hipertensión; Grupo 4: pacientes diabéticos con hipertensión postural. Se somete a cada persona a una tarea física y se registran dos variables en los tiempos 0, 1, 2, 3, 4, 5, 6, 8, 10, 12 y 15 minutos. Se desea determinar diferencia entre grupos.

1.3 - Se desea investigar la diferencia en la respuesta a dos drogas para el tratamiento de cierta enfermedad. Se utilizan los mismos sujetos para probar ambas drogas, por lo que hay dos fases (una para la aplicación de cada tratamiento) con un período intermedio para eliminar el efecto residual de la droga. En cada fase se registran los niveles de antibiótico en sangre en los tiempos 1, 2, 3 y 6 horas luego de la administración de la medicación.

1.4 - Se quiere estudiar la diferencia en la agudeza visual de sujetos con lentes de diferentes potencias. Para esto se registró el tiempo de respuesta al estímulo visual para los ojos izquierdo y derecho. Aquí las medidas repetidas no son temporales.

1.5 - Se desea estudiar la diferencia entre las respuestas de hombres y mujeres y el efecto edad, a una operación de cadera. Para ello se registró la edad y el sexo de los pacientes y se midieron los hematócritos una vez antes y tres veces después de la operación.

1.6 - Se desea estudiar el progreso en las habilidades para el álgebra de estudiantes a través de tres meses de instrucción. Un test estandarizado de álgebra es suministrado luego de 1 mes (nivel 1 del factor repetido Tiempo), y luego tests comparables son suministrados a los 2 y 3 meses (niveles 2 y 3 del factor tiempo respectivamente), registrándose los puntajes en los mismos.

En general, en este trabajo se usará el término "tiempo" para hacer referencia al factor repetido, y "grupo" para el factor cuya diferencia quiere analizarse.

Formas convencionales de analizar datos con medidas repetidas

Para analizar datos provenientes de ensayos con medidas repetidas pueden realizarse inicialmente gráficos (de fácil interpretación) con el fin de visualizar el comportamiento global de los datos. Sin embargo sólo con estos gráficos no puede realizarse una cuantificación de dicho comportamiento, ni compararse distintos grupos de unidades experimentales. Los métodos estadísticos utilizados con datos provenientes de mediciones repetidas apuntan a lograr esto último.

Los problemas de medidas repetidas son, sin lugar a dudas, multivariados pero hay una sola variable en estudio observada en distintas ocasiones en vez de varias variables medidas una sola vez cada una, como en el análisis multivariado clásico. Además, generalmente las observaciones se realizan en ocasiones seleccionadas en un continuo subyacente (tiempo), lo cual tampoco es lo habitual en el análisis multivariado. Por otra parte la natural secuencia de las observaciones lleva, en general, a tipos particulares de estructura de covarianza, situación que no se presenta en los datos multivariados donde no hay indicación de estructura alguna.

La posible dependencia de los datos provenientes de mediciones repetidas, introduce complicaciones adicionales al análisis. Para confiar en las conclusiones dadas por éstos debemos tener en cuenta esta posible dependencia. A pesar de ésto, la simplicidad de los métodos que asumen la independencia es tal que se han ideado alternativas para que puedan ser usados en este contexto. En esta línea de acción se pueden mencionar los siguientes análisis: 1) un análisis para cada tiempo y 2) análisis de aspectos de la respuesta o análisis de perfiles.

Comparación en cada tiempo

Una manera clásica de analizar datos de medidas repetidas consiste en analizar cada tiempo por separado. Este análisis es atractivo por su simplicidad, aunque tiene algunas desventajas. La única manera de determinar el cambio en los efectos de los tratamientos con el tiempo es comparando los análisis realizados separadamente. Una comparación similar podría haberse realizado usando grupos de sujetos independientes en cada tiempo. En este caso realizamos inferencias sobre el cambio en los promedios y no sobre el promedio de los cambios (que es el que interesa). Con esto perdemos toda la información que queríamos lograr usando medidas repetidas. Más aún, los tests no son independientes dado que los datos provienen de las mismas unidades experimentales. Si al querer comparar tratamientos los valores de un grupo son significativamente más grandes que los de otro en varias ocasiones, ésta no es fuerte evidencia de la superioridad de dicho grupo sobre el otro como lo sería si los tests fueran independientes (además de que se han realizado varios tests). Por otra parte, a veces se quiere decidir en qué momento ocurre cierto efecto, pero esto no tiene sentido dada la continuidad natural de los cambios. En resumen, este método es generalmente inválido.

Análisis del aspecto de la respuesta

Es un método estadísticamente válido aunque tiene poca potencia. En general se pone el énfasis en los procesos de generación de datos para producir modelos que ajusten correctamente los mismos. Sin embargo a veces interesa un aspecto particular del cambio en la respuesta a través del tiempo (por ejemplo conocer en qué momento se alcanza el "pico" de respuesta a una

droga, el tiempo que tarda un organismo en volver a su estado basal, etc.). Esto es lo que se denomina aspecto de la respuesta, pues resume ésta a lo largo del tiempo, y el análisis basado sólo en ella es el que nos interesa. Una ventaja importante de este método es que puede utilizarse aún cuando se tienen diferentes cantidades de mediciones y también cuando se han realizado en distintos tiempos (lo cual es muy común). El método pone su atención fundamentalmente en la forma de las curvas individuales, y no en la curva que pasa sobre la media en cada tiempo.

La principal desventaja de este método es la pérdida de grados de libertad del error. Si hay n individuos se calculan n estadísticos de resumen (uno por cada individuo) y los tests para comparar grupos deben estar basados en esos n valores derivados. Calculando un estadístico de resumen para cada sujeto, se obtiene un estimador preciso para cada uno, pero la variación *entre* sujetos no es considerada. Aquí hay algo importante implícito: aunque p (número de observaciones por individuo) sea grande, no compensa valores de n (tamaños de muestra) pequeños. La variación entre individuos sólo puede estimarse para tamaños de muestra suficientemente grandes. Para más detalle ver Frison y Pocock (1992)

Aunque pueden manipularse datos provenientes de muestras con distinta cantidad de mediciones en cada sujeto (datos censurados o perdidos), las varianzas de los mismos tienden a ser diferentes, por lo que se han ideado modificaciones a este análisis para tener en cuenta estos casos (sacrificando la simplicidad del análisis). Un tratamiento más exhaustivo puede verse en Hand y Crowder (1996).

Para probar los efectos que involucran los factores con medidas repetidas, se realiza el análisis de la varianza entre-grupos usual. Si el factor que corresponde a las medidas repetidas tiene más de dos niveles, hay dos alternativas para probar su significación: el análisis univariado (que es la manera usual de tratar estos datos) y el análisis multivariado (que requiere supuestos menos restrictivos). El enfoque multivariado permite la adopción de un modelo general para representar la estructura de covarianza de las observaciones. Dos desventajas de este tipo de análisis son la necesidad de observaciones completas y la baja potencia de los tests. El enfoque

univariando, en cambio, produce pruebas más sensibles, aunque requiere modelos más restrictivos para la estructura de covarianza de las observaciones. En la práctica ambos análisis llevan a resultados similares, a menos que los cambios a través de los niveles estén correlacionados con cambios a través de los sujetos. En el Ejemplo 1.6, se presenta el caso en que los progresos de los estudiantes del mes 1 al mes 2 están correlacionados con sus progresos del mes 2 al mes 3.

Consideraciones generales

Como es usual en estadística, diferentes modelos pueden generalmente ser aplicados a una amplia gama de experimentos.

Hay dos factores que proveen unicidad en los temas:

(1) los dos tipos de dependencia estocástica en las mediciones de la misma unidad experimental:
- homogeneidad en las respuestas de una unidad y heterogeneidad entre unidades;
- distancia (en tiempo o espacio) entre respuestas en una unidad.

(2) los tres tipos básicos de respuesta que pueden producirse:
- datos continuos;
- datos categóricos o de conteo;
- datos de duración o supervivencia.

A pesar de que las respuestas a los ensayos con medidas repetidas pueden ser de cualquiera de los tres tipos citados anteriormente, cabe destacar que nos centraremos en el caso continuo y en particular en aquellos estudios con medidas repetidas que contemplan el caso en que la distribución del error es normal.

El caso de normalidad es seleccionado porque:

. es adecuado para muchas situaciones reales;

. existen numerosas teorías, métodos y modelos desarrollados para tratarlos;

. la mayoría de los programas estadísticos tienen incorporadas rutinas para el análisis de los mismos.

Como se señaló, la presente monografía tratará de realizar una presentación de los métodos de análisis de la varianza (univariado y multivariado) y de regresión aplicados a datos provenientes de ensayos con medidas repetidas, cuando la distribución del error es normal. Además, como el interés está centrado en presentar un texto accesible para consulta de investigadores usuarios de la estadística, se proveen problemas resueltos (y la manera de resolverlos) utilizando algunos paquetes estadísticos de uso masivo, sin dejar de lado la interpretación de los resultados brindados por los mismos.

Bajo normalidad del error, los modelos de efectos aleatorios y modelos con matriz de covarianza con cierta estructura son abordados en detalle en Hand y Crowder (1996). Además el texto presenta el método general de estimación Gaussiana cuando la distribución del error no es normal, y trabaja con modelos no lineales, modelos lineales generalizados y estimadores de cuasi verosimilitud para datos binarios y categóricos.

Cuando las observaciones son independientes los modelos lineales generalizados (GLM) (Mc Cullagh y Nelder, 1989) y los de cuasi verosimilitud (Wedderburn, 1974; Mc Cullagh, 1983) se pueden aplicar para analizar una amplia gama de variables respuesta, tanto discretas como continuas. El problema reside en que con estos métodos no se tienen en cuenta posibles dependencias que pueden presentarse en las mediciones repetidas. Para estas situaciones se pueden aplicar las "ecuaciones de estimación generalizadas", conocidas como GEE. Este método propuesto por Liang y Zeger (1986) modela la variable respuesta incorporando la estructura de correlación existente entre las mediciones efectuadas, pudiendo además ser ellas de cualquier tipo (discretas o continuas).

CAPITULO 2
Análisis de la Varianza Multivariado

El presentar este análisis antes del univariado se debe a que es conceptualmente más simple de aplicar, dado que no hay supuestos sobre la forma de la matriz de covarianza de los errores y tiene la ventaje de proporcionar una gran facilidad de interpretación. El enfoque multivariado es descripto en detalle en Hand y Taylor (1987).

En el contexto de medidas repetidas decir "análisis multivariado" es equivalente a decir que ninguna estructura particular se asume para la matriz de covarianza de los errores. Aunque en cierta manera éste es un análisis robusto, puede dar por resultado inferencias débiles, en el sentido que se usan muchos grados de libertad para estimar los parámetros de covarianza dejando muy pocos para los parámetros de interés.

Ideas básicas

Sea y_{ij} la j-ésima medición del i-ésimo sujeto. Las p observaciones del sujeto i se agrupan en el vector $\boldsymbol{y}_i$ y si hay n sujetos, podemos reagrupar los vectores en una matriz nxp como la siguiente:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \cdots \\ \boldsymbol{y}_n \end{pmatrix} = \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ . & . & . \\ y_{n1} & \cdots & y_{np} \end{pmatrix}$$

En general los $\boldsymbol{y}_i$ provienen de distintos grupos (pueden ser estructuras de tratamientos con individuos asignados aleatoriamente a distintas condiciones), y pueden modelarse las observaciones en función de las medias de los grupos y de los desvíos aleatorios de esas medias:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Xi} + \boldsymbol{U}$$

donde $\boldsymbol{X}$ es una matriz nxq (q cantidad de variables explicativas o de grupos) que puede incluir covariables que no cambien con el tiempo. Se la llama *matriz de diseño, matriz entre individuos* o *matriz entre grupos*, ya que su función es distinguir entre los individuos de los distintos grupos y sus filas se "aplican" a todas las ocasiones. Por su parte Ξ contiene las respuestas medias en cada grupo.

En los ejemplos que se dan más adelante se mostrará la forma de ésta y de las matrices $\boldsymbol{X}$ y Ξ para distintos casos. En general, la i-ésima fila de $\boldsymbol{X}\Xi$ contiene la respuesta media para el sujeto i, y la i-ésima fila de $\boldsymbol{U}$ los desvíos aleatorios respecto de las medias; es decir, las filas de $\boldsymbol{X}\Xi$ dan las respuestas medias que serían observadas si el estudio se repitiera muchas veces para cada individuo, y las filas de $\boldsymbol{U}$ cuánto se alejan los datos del presente estudio de dichas medias. Cabe destacar que en $\boldsymbol{U}$ están incluídas todas las variaciones aleatorias, y que, dado que sus filas corresponden a diferentes sujetos las mismas son independientes. Como asumimos que las distribuciones de los errores son normales multivariadas, la i-ésima fila de $\boldsymbol{U}$ se distribuye como una normal multivariada $N_p(0, \boldsymbol{\Sigma}_p)$. De esta manera resulta que la respuesta media es

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\Xi \qquad \textbf{(2.1)}$$

Si $p = 1$ (es decir hay una observación para cada individuo), éste es el modelo lineal general univariado con distinto conjunto de parámetros para cada grupo.

Ejemplo 2.1: En una situación como la del Ejemplo 1.6 del capítulo anterior, se tienen 3 sujetos, a cada uno de los cuales se los sometió a pruebas estandarizadas de álgebra luego de 1, 2 y 3 meses de recibir instrucción, registrándose el puntaje en las mismas. Los datos obtenidos fueron:

$$\boldsymbol{Y} = \begin{pmatrix} 24 & 26 & 28 \\ 30 & 30 & 32 \\ 30 & 29 & 28 \end{pmatrix}$$

donde $y_{11} = 24$ indica que el Sujeto 1 en la Prueba 1 obtuvo 24 puntos,

$y_{12} = 26$ indica que el Sujeto 1 en la Prueba 2 obtuvo 26 puntos, etc.

Es decir, en la fila i-ésima figuran los valores correspondientes al sujeto i. Para estos datos tenemos

$$\boldsymbol{X} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad\qquad \Xi = (\,\xi_1 \quad \xi_2 \quad \xi_3\,)$$

con ξ_i la calificación media de todos los sujetos en la Prueba $i\,(i = 1, 2, 3)$.

Ejemplo 2.2: Si en el ejemplo anterior tenemos 5 individuos $(n = 5)$ en 2 grupos (el primero de 3 sujetos y el segundo de 2) determinados por dos tipos de instrucción, a los cuales observamos 1 vez $(p = 1)$, la ecuación (2.1) queda

$$\mathrm{E}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$$

con ξ_1 la calificación media del grupo 1 y ξ_2 la calificación media del grupo 2.

Ejemplo 2.3: Una parametrización alternativa a la usada en el ejemplo anterior es:

$$\mathrm{E}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$$

ahora con ξ_1 la calificación media del grupo 1 y ξ_2 la diferencia entre las calificaciones medias del grupo 1 y el grupo 2.

Ejemplo 2.4: Si ahora las mediciones se realizaran en 3 tiempos $(p = 3)$ es decir se sometiera a los alumnos a pruebas luego de 1, 2 y 3 meses de instrucción, resulta que:

$$\mathsf{E}\begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \\ y_{41} & y_{42} & y_{43} \\ y_{51} & y_{52} & y_{53} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_{11} & \xi_{12} & \xi_{13} \\ \xi_{21} & \xi_{22} & \xi_{23} \end{pmatrix}$$

donde cada columna de la última matriz se corresponde con las observaciones de un tiempo determinado. Aquí, por ejemplo, ξ_{12} es la calificación media del Grupo 1 luego de 2 meses de instrucción, mientras ξ_{21} es la calificación media del Grupo 2 luego de 1 mes de instrucción. En este caso hacer un análisis multivariado parece ser lo más adecuado.

Ejemplo 2.5: Supongamos que en el ejemplo anterior, la edad al comenzar el estudio es relevante. Sea x_i la edad del individuo i, entonces

$$\mathsf{E}\begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \\ y_{41} & y_{42} & y_{43} \\ y_{51} & y_{52} & y_{53} \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 1 & 0 & x_3 \\ 0 & 1 & x_4 \\ 0 & 1 & x_5 \end{pmatrix} \begin{pmatrix} \xi_{11} & \xi_{12} & \xi_{13} \\ \xi_{21} & \xi_{22} & \xi_{23} \\ \xi_{31} & \xi_{32} & \xi_{33} \end{pmatrix}$$

donde ξ_{3j} es el coeficiente para la edad en la j-ésima ocasión, por lo que el "efecto edad" sobre el i-ésimo sujeto en la j-ésima ocasión es $x_i\xi_{3j}$.

El modelo (2.1) es el clásico de análisis de la varianza multivariado (MANOVA) que define una estructura entre-sujetos para el valor esperado de las observaciones, pero no define ninguna relación entre variables. Como esto último es lo que queremos en general (imponer restricciones sobre las observaciones en cada ocasión), asumiremos que la matriz Ξ proviene del modelo:

$$\Xi = \Gamma B$$

donde Γ es una matriz qxr (que en los ejemplos que se verán a continuación contiene los parámetros de regresión para cada grupo) y B es la matriz rxp que describe el modelo para el patrón de cambio a través del tiempo (perfil de respuesta esperada dentro de un sujeto). Por esto a B se la llama *matriz dentro de sujeto* o *matriz dentro de grupo*. De aquí resulta que

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X\Gamma B} \qquad \textbf{(2.2)}$$

Ejemplo 2.6: Supongamos tener un solo grupo de 3 sujetos, y que en el Ejemplo 2.1 creemos que la respuesta está linealmente relacionada al tiempo. Allí medimos la respuesta en 3 tiempos (t_1, t_2, t_3). De acuerdo al nuevo modelo la respuesta media en el tiempo t es $\gamma_1 + \gamma_2 t$. Esto puede escribirse como:

$$\mathrm{E}\begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} \gamma_1 & \gamma_2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ t_1 & t_2 & t_3 \end{pmatrix}$$

con t_j la covariable que toma diferentes valores en distintos tiempos.

Ejemplo 2.7: En similares condiciones que en el ejemplo anterior pero con dos grupos (uno de tres y el otro de dos individuos) como en el Ejemplo 2.4, resulta que:

$$\mathrm{E}\begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \\ y_{41} & y_{42} & y_{43} \\ y_{51} & y_{52} & y_{53} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ t_1 & t_2 & t_3 \end{pmatrix}$$

Aunque aquí hay una relación lineal con el tiempo dentro de cada grupo, las pendientes y las ordenadas al origen pueden ser diferentes entre sí. Los individuos del grupo 1 tienen regresión $\gamma_{11} + \gamma_{12} t$ mientras que los del grupo 2 tienen $\gamma_{21} + \gamma_{22} t$.

También pueden incluirse en el análisis covariables que cambien con el tiempo con otra parametrización más conveniente o más relacionada con las hipótesis investigadas.

Ejemplo 2.8: Una parametrización alternativa para el ejemplo anterior es la siguiente:

$$\mathsf{E}\begin{pmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \\ y_{31} & y_{32} & y_{33} \\ y_{41} & y_{42} & y_{43} \\ y_{51} & y_{52} & y_{53} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$$

Aquí en vez de expresar la respuesta como una función lineal del tiempo, la describimos en términos de la desviación con respecto a la media general μ_{k1} para el grupo k. Así, las medias del grupo k en las tres ocasiones son $\mu_{k1} + \mu_{k2}$, $\mu_{k1} + \mu_{k3}$ y $\mu_{k1} - \mu_{k2} - \mu_{k3}$. La tercera está construída para que el promedio de las tres medias sea igual a μ_{k1} .

Estimadores de los parámetros

Los estimadores de máxima verosimilitud de los parámetros del modelo (2.2), según puede verse en Khatri (1966), están dados por

$$\hat{\boldsymbol{\Gamma}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\hat{\boldsymbol{\Sigma}}_p^{-1}\boldsymbol{B}^T(\boldsymbol{B}\hat{\boldsymbol{\Sigma}}_p^{-1}\boldsymbol{B}^T)^{-1} \qquad \textbf{(2.3)}$$

donde $\hat{\boldsymbol{\Sigma}}_p$ es el estimador de máxima verosimilitud usual de la matriz de covarianza de las filas de $\boldsymbol{U}$ (como siempre para que $\boldsymbol{B}\hat{\boldsymbol{\Sigma}}_p^{-1}\boldsymbol{B}^T$ sea no singular se requiere que el rango de $\boldsymbol{B}$ sea r, y para que $\boldsymbol{X}^T\boldsymbol{X}$ sea no singular se requiere que el rango de $\boldsymbol{X}$ sea q).

En particular, si hubiera una sola observación para cada individuo, entonces

$$\hat{\boldsymbol{\Gamma}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

que es la solución usual de mínimos cuadrados para el modelo $\mathsf{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\Gamma}$.

Similarmente si suponemos que hay un solo individuo medido en p ocasiones, el estimador es

$$\hat{\boldsymbol{\Gamma}} = \boldsymbol{Y}\hat{\boldsymbol{\Sigma}}_p^{-1}\boldsymbol{B}^T(\boldsymbol{B}\hat{\boldsymbol{\Sigma}}_p^{-1}\boldsymbol{B}^T)^{-1}$$

por ser $\boldsymbol{X} = (1)$.

Ésta es la solución por mínimos cuadrados ponderados teniendo en cuenta las correlaciones entre los resultados obtenidos en los distintos tiempos. Y juntando las dos soluciones resulta (2.3). Así ambas estimaciones coinciden.

Variables derivadas

El análisis del apecto de la respuesta puede, según Hand y Crowder (1996), extenderse a más de una característica y resumir, por ejemplo, cambios a través del tiempo de tipo lineal, cuadrático, cúbico, etc.. Si el interés radica en estudiar conjúntamente esos cambios se deben generar nuevas varibles a partir de las medidas que son llamadas en este contexto variables derivadas.

Los estimadores de máxima verosimilitud por sí solos pueden ser usados para realizar tests de razón de verosimilitud con el fin de probar la adecuación del modelo, pero usualmente se generan variables derivadas a partir del siguiente procedimiento. Las filas de la matriz $\boldsymbol{\Gamma B}$ contienen las combinaciones lineales de los vectores fila de $\boldsymbol{B}$, es decir, dada una fila de $\boldsymbol{\Gamma}$, el producto da como resultado un punto en el espacio fila de $\boldsymbol{B}$. Una manera equivalente de describir estos puntos es cambiando la base del espacio fila de $\boldsymbol{B}$ postmultiplicando $\boldsymbol{\Gamma B}$ por una matriz $\boldsymbol{H}_1$ (de orden pxr) cuyas columnas expanden el espacio fila de $\boldsymbol{B}$. Las p filas componentes de E($\boldsymbol{Y}$) yacen en un subespacio r-dimensional del espacio p-dimensional. Postmultiplicando por $\boldsymbol{H}_1$ la transformamos en E($\boldsymbol{Y}$)$\boldsymbol{H}_1$ con r filas componentes. Si expandimos $\boldsymbol{H}_1$ a una matriz pxp $\boldsymbol{H} = (\boldsymbol{H}_1, \boldsymbol{H}_2)$ de rango completo, donde las columnas de $\boldsymbol{H}_2$ son ortogonales a las de $\boldsymbol{H}_1$ (y por ende a las filas de $\boldsymbol{B}$), entonces definiendo $\boldsymbol{Y}_1 = \boldsymbol{Y}\boldsymbol{H}_1$ e $\boldsymbol{Y}_2 = \boldsymbol{Y}\boldsymbol{H}_2$ obtenemos:

$$\text{E}(\boldsymbol{Y}_1) = \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{B}\,\boldsymbol{H}_1 = \boldsymbol{X}\boldsymbol{\Xi}\boldsymbol{H}_1 = \boldsymbol{X}\boldsymbol{\Theta} \quad \text{y} \quad \text{E}(\boldsymbol{Y}_2) = \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{B}\,\boldsymbol{H}_2 = \boldsymbol{X}\boldsymbol{\Xi}\boldsymbol{H}_2 = \mathbf{0}$$

$\boldsymbol{H}_1$ es elegida de manera tal que ciertas columnas resuman el patrón de cambio a través del tiempo de forma útil. Por ejemplo, la primera columna podría contener los coeficientes de tal forma que la primera columna de $\boldsymbol{Y}\boldsymbol{H}_1$ contenga las medias de las p observaciones de cada individuo, la segunda podría tener la tendencia lineal, y así sucesivamente. Las columnas de $\boldsymbol{H}_2$

deben contener aquellas combinaciones lineales de las p observaciones de cada individuo que tengan esperanza cero.

Para probar si el modelo provee un ajuste adecuado a las medias pueden usarse los tests multivariados usuales para verificar que $\mathrm{E}(\boldsymbol{Y}_2) = \mathrm{E}(\boldsymbol{Y}\boldsymbol{H}_2) = \mathbf{0}$. Esto es, transformamos la matriz de observaciones $\boldsymbol{Y}$ por $\boldsymbol{H}_2$ y vemos que las medias derivadas son cero.

Ejemplo 2.9: Supongamos como en el Ejemplo 2.6 (con $t_1 = 1$, $t_2 = 2$ y $t_3 = 3$)

$$\boldsymbol{X} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad \boldsymbol{\Gamma} = \begin{pmatrix} \gamma_1 & \gamma_2 \end{pmatrix} \qquad \boldsymbol{B} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix}$$

Entonces

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{B}$$

$$= \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} \gamma_1 & \gamma_2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix}$$

$$= \begin{pmatrix} \gamma_1 + \gamma_2 & \gamma_1 + 2\gamma_2 & \gamma_1 + 3\gamma_2 \\ \gamma_1 + \gamma_2 & \gamma_1 + 2\gamma_2 & \gamma_1 + 3\gamma_2 \\ \gamma_1 + \gamma_2 & \gamma_1 + 2\gamma_2 & \gamma_1 + 3\gamma_2 \end{pmatrix}$$

y así el modelo es tal que hay un solo grupo con tres sujetos, con la respuesta linealmente relacionada al tiempo (con ordenada al origen γ_1 y coeficiente de regresión γ_2).

Ahora definimos

$$\boldsymbol{H}_1 = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \qquad \text{y} \qquad \boldsymbol{H}_2 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

La manera de elegir $\boldsymbol{H}_1$ no es única; más adelante se verán otras opciones.

Para comprobar que las elecciones anteriores son válidas, primero veremos que efectivamente las columnas de $\boldsymbol{H}_1$ expanden el espacio fila de $\boldsymbol{B}$. Tenemos la primera fila de $\boldsymbol{B}$

$$(1 \quad 1 \quad 1) = 1.(1 \quad 1 \quad 1) + 0.(-1 \quad 0 \quad 1)$$

donde los vectores fila del lado derecho son las columnas de $\boldsymbol{H}_1$ y los coeficientes se obtienen a partir de conceptos de espacios vectoriales.

Similarmente, para la segunda fila de $\boldsymbol{B}$,

$$(1 \quad 2 \quad 3) = 2.(1 \quad 1 \quad 1) + 1.(-1 \quad 0 \quad 1)$$

Y las columnas de $\boldsymbol{H}_2$ son ortogonales a las de $\boldsymbol{H}_1$:

$$(1 \quad -2 \quad 1)\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 0 \qquad \text{y} \qquad (1 \quad -2 \quad 1)\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} = 0$$

Ahora podemos probar la adecuación del modelo (es decir que la respuesta crece linealmente con el tiempo) comparando los apartamientos de la linealidad con cero, esto es comparando $\boldsymbol{X\Gamma BH}_2$ con cero. Esta expresión es la E($\boldsymbol{Y}_2$) o E($\boldsymbol{YH}_2$), por lo que estamos transformando las filas de $\boldsymbol{Y}$ postmultiplicando por $\boldsymbol{H}_2$. En este caso, dado que $\boldsymbol{H}_2$ es un vector columna la transformación da por resultado un número para cada sujeto. Para ver si la media de esos valores es cero puede usarse un test t para una muestra.

En realidad el objetivo es modelar E($\boldsymbol{X\Gamma B}$), pero dado que ajustamos ese modelo a través de una transformación de las observaciones, deben estudiarse las transformaciones con mayor profundidad.

Ejemplo 2.10: Supongamos tener dos grupos de sujetos medidos cada uno en tres tiempos. Hay entonces tres preguntas fundamentales que deben tenerse en cuenta:

1 - ¿ Tienen los perfiles de las medias de los grupos el mismo nivel ?. Es decir, ¿hay "efecto grupo" ?.

2 - ¿ Son los perfiles planos ?. Es decir, ¿ hay "efecto tiempo" ?

3 - ¿ Son los perfiles paralelos ?. Es decir, ¿hay "interacción grupo por tiempo"?

En la práctica se debería responder primero a la pregunta 3, pues de existir evidencias de interacción no tendría sentido analizar las otras dos. Ahora responderemos estas cuestiones en el orden en que se enuncian para simplificar este primer análisis.

Para definir variables derivadas que permitan responder a las preguntas planteadas podemos postmultiplicar la matriz de datos por

$$\boldsymbol{H} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix}$$

Consideremos los resultados para un sujeto particular $\boldsymbol{y} = (y_1 \quad y_2 \quad y_3)$. Entonces:

$$\begin{aligned} \boldsymbol{yH} &= (y_1 \quad y_2 \quad y_3) \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} \\ &= (y_1 + y_2 + y_3 \qquad y_1 - y_2 \qquad y_2 - y_3) \\ &= (z_1 \quad z_2 \quad z_3) \end{aligned}$$

La primera variable derivada z_1 es proporcional a la media de los tres resultados. La segunda y tercera conjuntamente resumen las posibles diferencias entre los tres resultados.

La pregunta 1 concierne a las medias de los datos; en términos de las variables derivadas, la hipótesis nula es

$$\mathrm{H}_0 : \mathrm{E}(z_1)_{Grupo\,1} = \mathrm{E}(z_1)_{Grupo\,2}.$$

Como en este caso hay sólo dos grupos podemos usar para probar esta hipótesis un test t.

La pregunta 2, tiene relación con la "forma" de los perfiles (patrones de diferencia entre y_1, y_2 e y_3). Las variables derivadas z_2 y z_3 resumen estas diferencias: si ambas son cero en los dos grupos, no hay efecto tiempo. Por lo tanto para responder a esta pregunta planteamos:

$$H_0 : E(z_2 , z_3) = (0, 0)$$

Esta es una hipótesis nula bivariada que puede probarse usando el test T^2 de Hotelling (extensión natural del test t, que se verá más adelante).

Por último para la pregunta 3 que trata sobre si el patrón de cambio es el mismo en los dos grupos, nuevamente las variables derivadas z_2 y z_3 son las que deben utilizarse. Aquí la hipótesis es

$$H_0 : E(z_2 , z_3)_{Grupo\,1} = E(z_2 , z_3)_{Grupo\,2}$$

que también puede analizarse utilizando el test T^2 de Hotelling .

En el Ejemplo 2.10 se definieron las variables derivadas z_2 y z_3 usando $(1 \quad -1 \quad 0)$ y $(0 \quad 1 \quad -1)$. En conjunto estas variables expanden el espacio de posibles diferencias. Otras variables derivadas pueden usarse con el mismo propósito; por ejemplo, $(1 \quad 0 \quad -1)$ y $(1 \quad -2 \quad 1)$.

Ambos conjuntos son equivalentes dado que:

$$(1 \quad 0 \quad -1) = (1 \quad -1 \quad 0) + (0 \quad 1 \quad -1)$$

y

$$(1 \quad -2 \quad 1) = (1 \quad -1 \quad 0) - (0 \quad 1 \quad -1)$$

En general si $\boldsymbol{H}$ es uno de tales conjuntos resulta que $\boldsymbol{HG}$ es otro de esos conjuntos, con $\boldsymbol{G}$ una matriz invertible de orden pxp .La elección de uno u otro depende del interés del investigador; por ejemplo el conjunto $(1 \quad 0 \quad -1)$ y $(1 \quad -2 \quad 1)$ contiene los coeficientes de polinomios ortogonales correspondientes a los términos lineal y cuadrático, respectivamente, los cuales resultan útiles para describir el cambio de la respuesta a través del tiempo, como ya fuera comentado.

La solución por máxima verosimilitud de la ecuación (2.2) presentada en (2.3) es la solución por mínimos cuadrados generalizados, en la que las correlaciones entre los tiempos es considerada.

Usando el Lema 1 de Khatri (1966),

$$\hat{\Gamma} = (X^T X)^{-1} X^T Y R B^T (B B^T)^{-1}$$

donde YR son los residuos luego de hacer la covarianza de YH_2 con H_2 una matriz $pxp-r$ que satisface que $BH_2 = 0$ y $H_2^T H_2 = I$, es decir

$$YR = Y(I - P_2) = Y\{ I - H_2(H_2^T \Sigma_p H_2)^{-1} H_2^T \Sigma_p \}$$

Así el estimador de máxima verosimilitud es equivalente a un simple estimador mínimos cuadrados basado en las variables derivadas formadas al realizar las covarianzas de aquellas variables que tienen esperanza cero bajo el modelo. Esta equivalencia sugiere una generalización en la cual algunas (no todas) las variables derivadas deben ser "covariadas".

Estructuras de medidas repetidas más complejas

Generalmente aparecen estructuras de mediciones repetidas más complejas que las presentadas. Por ejemplo, *medidas repetidas múltiples*, en la que varias variables son medidas en cada ocasión y hay relación entre las mismas. Un caso en que esto sucede es cuando se aplican, por ejemplo, 5 drogas o dosis a cada individuo y se mide la respuesta a la misma en 4 tiempos diferentes; en ese caso se tiene un problema de medidas repetidas doble. También es común que se presenten estos problemas en áreas como psicología donde se miden respuestas en intervalos pequeños de tiempo (minutos) y luego el experimento completo se repite en períodos más prolongados (días). La situación 1.4 antes presentada es otro de estos casos.

La matriz H sirve para re-expresar las variables fila en términos de variables derivadas que representan efectos principales e interacciones entre y dentro de los sujetos.

Ejemplo 2.11: Si son tomadas seis mediciones a cada sujeto, las primeras tres luego de 1, 2 y 3 horas de administrar la dosis 1 de cierta droga, y las restantes luego de 1, 2 y 3 horas de administrar la dosis 2 de la misma droga, la matriz H resultante es:

$$\boldsymbol{H} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & -2 & 0 & -2 \\ 1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & -2 & 0 & 2 \\ 1 & -1 & -1 & 1 & 1 & -1 \end{pmatrix}$$

donde la columna 1 da el nivel de respuesta medio general, la 2 la variable derivada para medir el "efecto dosis", la 3 y la 4 las que miden el "efecto tiempo" (la 3 el efecto lineal y la 4 el cuadrático), y la 5 y 6 resumen la interacción dosis por tiempo: la Columna 5 es $2x$Columna 3 (componente lineal de la interacción), y la Columna 6 es $2x$Columna 4 (componente cuadrática).

En general los programas estadísticos agrupan las variables derivadas de acuerdo al orden natural y luego realizan tests separados para los efectos principales y las interacciones.

Tests multivariados

Ya se ha mostrado la manera de transformar el vector de observaciones a través de la matriz $\boldsymbol{H}$ en un conjunto de variables derivadas que sirven para probar ciertas cuestiones de interés. En el Ejemplo 2.10 se mostraron dos conjuntos de variables: el primero (z_1) relativo a la media de las respuestas, y el segundo (z_2, z_3) que se usa para probar tendencia con el tiempo. Si una de las cuestiones de interés (por ejemplo la relacionada a la media) es respondida usando una sola variable derivada, entonces el análisis que se usa es un análisis de la varianza univariado. Si, en cambio, debe usarse más de una de las variables derivadas para responder una de las preguntas se debe realizar un test multivariado.

En el caso univariado se trabaja con estadísticos con distribución F, que son razones entre las sumas de cuadrados de la hipótesis y la residual. En los tests multivariados se realiza una extensión del caso univariado, reemplazando las sumas de cuadrados por matrices, en cuyas diagonales están las sumas de cuadrados correspondientes a cada una de las p variables y fuera de la diagonal los términos de productos cruzados; además se reemplaza la razón

de las sumas de cuadrados por el producto de la matriz de la hipótesis y la inversa de la matriz del error.

Para comprender estos tests comenzamos con el caso en que hay una sola variable y, distribuida como N(μ, σ^2) y medida a un solo grupo de n individuos. La hipótesis de interés es

$$H_0: \mu = \mu_0$$

y el estadístico es el conocido

$$t = \frac{\overline{y}-\mu_0}{s/\sqrt{n}}$$

el cual puede ser re-escrito como

$$t^2 = (\overline{y} - \mu_0)\, s^{-2} (\overline{y} - \mu_0)\, n \qquad \textbf{(2.4)}$$

Ahora, sea $\boldsymbol{y}$ un vector p variado con vector de medias $\boldsymbol{\mu}$ para el que se postula

$$H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$$

Esto puede reducirse al caso univariado mediante una combinación lineal de sus componentes, $\mathbf{a}^T\boldsymbol{y}$, a la que luego se le aplicará un test t.

Ahora, si $\mathbf{a}^T\boldsymbol{\mu} = \mathbf{a}^T\boldsymbol{\mu}_0$ para todas las combinaciones no nulas **a**, entonces resulta $\boldsymbol{\mu} = \boldsymbol{\mu}_0$. Esto parece en principio difícil de probar, pero se puede ver que si **a** es aquel que brinda el valor más grande de t que no lleva a rechazar la hipótesis $\mathbf{a}^T\boldsymbol{\mu} = \mathbf{a}^T\boldsymbol{\mu}_0$ no hay otro valor de **a** que lleve a rechazarla (principio de "unión-intersección").

La media muestral de $\mathbf{a}^T\boldsymbol{y}_i$ es $\mathbf{a}^T\overline{\boldsymbol{y}}$ que, bajo la hipótesis nula, tiene media $\mathbf{a}^T\boldsymbol{\mu}_0$ y desviación estándar estimada $\sqrt{\mathbf{a}^T\boldsymbol{S}_p\mathbf{a}}$, por lo que el estadístico es:

$$t(\mathbf{a}) = \frac{\mathbf{a}^T\overline{\boldsymbol{y}}-\mathbf{a}^T\boldsymbol{\mu}_0}{\sqrt{\mathbf{a}^T\boldsymbol{S}_p\mathbf{a}}/\sqrt{n}}$$

Deseamos encontrar el **a** que maximice la expresión anterior, aunque ésto no determina unívocamente a **a**, ya que multiplicando por una constante arbitraria llegamos a la misma solución. Para evitar esta situación se impone que $\mathbf{a}^T \boldsymbol{S}_p \mathbf{a} = 1$ y se maximiza $t(\mathbf{a})$ con esa condición.

Usando multiplicadores de Lagrange se llega a que el $max_{\mathbf{a}}\, t^2(\mathbf{a})$ es un solo valor no nulo $\boldsymbol{\lambda}$ tal que

$$|(\overline{\boldsymbol{y}} - \boldsymbol{\mu}_0)(\overline{\boldsymbol{y}} - \boldsymbol{\mu}_0)^T n - \boldsymbol{\lambda} \boldsymbol{S}_p| = 0$$

el cual es

$$T^2 = (\overline{\boldsymbol{y}} - \boldsymbol{\mu}_0)^T \boldsymbol{S}_p^{-1} (\overline{\boldsymbol{y}} - \boldsymbol{\mu}_0) n$$

que es una versión multivariada de (2.4). Este es el **estadístico T^2de Hotelling**. Bajo la hipótesis nula,

$$\frac{n-p}{p(n-1)} T^2 \sim F(p, n-p)$$

Para el caso de dos grupos,

$$\mathrm{H}_0\colon \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

y la solución está dada por la raíz de

$$\left|(\overline{\boldsymbol{y}}_1 - \overline{\boldsymbol{y}}_2)(\overline{\boldsymbol{y}}_1 - \overline{\boldsymbol{y}}_2)^T \frac{n_1 n_2}{n_1+n_2} - \boldsymbol{\lambda} \boldsymbol{S}_p\right| = 0$$

que da por resultado

$$T^2 = (\overline{\boldsymbol{y}}_1 - \overline{\boldsymbol{y}}_2)^T \boldsymbol{S}_p^{-1} (\overline{\boldsymbol{y}}_1 - \overline{\boldsymbol{y}}_2) \frac{n_1 n_2}{n_1+n_2}$$

con

$$\frac{n_1+n_2-p-1}{(n_1+n_2-2)p} T^2 \sim F(p, n_1 + n_2 - p - 1)$$

Aquí las observaciones del Grupo 1 se asumen como $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_p)$ y las del Grupo 2 como $\mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_p)$. El supuesto de igualdad de matrices de covarianza es riesgoso, pero se sabe que con tamaños de muestra grandes e iguales, la

desigualdad de las matrices de covarianza tiene escaso efecto en el error de tipo I.

Los dos casos anteriores tienen soluciones dadas por las raíces de ecuaciones de la forma

$$|\boldsymbol{T} - \boldsymbol{\lambda E}| = 0$$

donde $\boldsymbol{E}$ es la matriz de las sumas de cuadrados del error dentro de grupo y $\boldsymbol{T}$ es la de las sumas de cuadrados de la hipótesis entre grupos. Esto permite una generalización a más de dos grupos.

Esta generalización a clases múltiples (tales que la matriz de la hipótesis tenga rango mayor que 1) introduce la complicación adicional de que hay más de una raíz para la ecuación presentada. Cada una de ellas (denotadas $\lambda_1,...,\lambda_r$) representa un componente de la variación entre grupos. Distintos estadísticos se definen combinando esas *raíces características* o *autovalores*.

Por ejemplo:

Estadístico del autovalor más grande de Roy: es λ_1 que es la raíz más grande. Generalmente no se define como el autovalor más grande de $\boldsymbol{TE}^{-1}$ sino (equivalentemente) como el autovalor mayor de $\boldsymbol{T}(\boldsymbol{T}+\boldsymbol{E})^{-1}$.

Traza de Hotelling-Lawley: es $\Sigma\,\boldsymbol{\lambda}_i$.

Estadístico del test de razón de verosimilitud* o *Lambda de Wilks: definido como $\Pi(1+\boldsymbol{\lambda}_i)^{-1}$.

Traza de Pillai-Bartlett: es $\Sigma\,\boldsymbol{\lambda}_i/(1+\boldsymbol{\lambda}_i)$, que es la traza de $\boldsymbol{T}(\boldsymbol{T}+\boldsymbol{E})^{-1}$.

El estadístico del test univariado sigue una distribución F cuando la hipótesis nula es cierta. En el caso multivariado esto no es siempre así, aunque pueden aplicarse transformaciones para dar una aproximación a la distribución F. Éstas consisten principalmente en interpolaciones, lo cual explica la razón por la que los programas dan distribuciones F con grados de

libertad no enteros. Aproximaciones asintóticas de las distribuciones de estos estadísticos a través de distribuciones Chi-cuadrado o F son consideradas en Anderson (1958).

Muchos programas dan los valores de los cuatro estadísticos mencionados, para que luego se elija el más apropiado. Generalmente deseamos utilizar aquel que brinde el test más potente, aunque esto depende en gran medida de los apartamientos de la hipótesis nula de igualdad de medias de los vectores. Si la diferencia yace a lo largo de un "continuo univariado" el estadístico de Roy parece ser el más potente, y la potencia disminuye a medida que nos movemos hacia abajo en la lista dada de estadísticos. Sin embargo en ciertos casos las potencias siguen el orden inverso; algunos autores prefieren el de Pillai-Bartlett, el cual es el más robusto a la presencia de "outliers".

Cuando sólo se comparan dos grupos, la matriz $\boldsymbol{TE}^{-1}$ tiene rango 1, y los estadísticos anteriores dan todos el mismo resultado.

Tests para los supuestos

Para realizar el análisis de la varianza multivariado se supone que los datos provienen de una distribución normal multivariada, y que las distribuciones de cada clase (factores entre sujetos) tienen la misma matriz de covarianza. Este último supuesto es una extensión del de homogeneidad de varianza del análisis univariado y, como en aquel caso, las probabilidades de error son menos afectadas si el supuesto es falso cuando los tamaños de las muestras son aproximadamente iguales y no muy pequeños.

Una distribución normal multivariada necesariamente tiene distribuciones marginales normales, lo cual puede examinarse realizando histogramas, diagramas de cajas, de tallo-hoja o de probabilidades en cada ocasión separadamente. Esto puede extenderse a diagramas bivariados que permiten analizar pares de variables simultáneamente. Sin embargo si hay normalidad de las marginales, esto no implica normalidad multivariada. Existen tests para probar la normalidad multivariada y se debe tener cuidado al rechazar una

comparación robusta de medias usando un test muy sensible a los apartamientos de la normalidad.

Hay dos tests usuales para probar igualdad de matrices de covarianza, basados en el criterio de razón de verosimilitud.

La hipótesis nula es

$$H_0 : \boldsymbol{\Sigma}_{p1} = \boldsymbol{\Sigma}_{p2} = \ldots = \boldsymbol{\Sigma}_{pg}$$

donde $\boldsymbol{\Sigma}_{pi}$ denota la matriz de covarianza del Grupo i , versus

$$H_1 : \boldsymbol{\Sigma}_{pr} \neq \boldsymbol{\Sigma}_{ps} \text{ para algún } r \text{ y } s$$

Siendo $\boldsymbol{S}_{pr}$ el estimador insesgado usual de $\boldsymbol{\Sigma}_{pr}$ y

$$\boldsymbol{S}_p = \sum \left[(n_r - 1)\boldsymbol{S}_{pr} \right] / \sum (n_r - 1)$$

el estimador de la matriz de covarianza común, el estadístico del test de razón de verosimilitud es:

$$\boldsymbol{M} = \sum (n_r - 1) \ln |\boldsymbol{S}_p| - \sum (n_r - 1) \ln |\boldsymbol{S}_{pr}|$$

En principio $\boldsymbol{M}k$ sigue aproximadamente una distribución χ^2 con $(g-1)p(p+1)/2$ grados de libertad, con k definido como:

$$k = 1 - \frac{2p^2+3p-1}{6(p+1)(g-1)} \left(\sum_r \frac{1}{(n_r-1)} - \frac{1}{\Sigma(n_r-1)} \right)$$

Algunos programas estadísticos muestran estos resultados.

Para poner en claro los conceptos dados anteriormente se presenta a continuación una aplicación de esta prueba realizada con el software estadístico SAS, similar al Ejemplo 2.10 mencionado anteriormente en este capítulo.

Problema 1: Enfoque multivariado

Para investigar el efecto de un suplemento dietario de vitamina E en el crecimiento de cerdos "guinea" se llevó a cabo el siguiente estudio. Se tomaron 15 animales, se les suministró durante una semana una sustancia inhibidora del crecimiento, y la terapia con vitamina E se comenzó al principio de la semana 5. Cada uno de los 5 animales de cada grupo recibió respectivamente cero, baja y alta dosis de vitamina E y se registró el peso corporal de cada animal al final de las semanas 1, 3, 4, 5, 6 y 7.

El peso corporal (en gramos) de los animales se da en la siguiente tabla.

	Semana					
Grupo	1	3	4	5	6	7
1	455	460	510	504	436	466
1	467	565	610	596	542	587
1	445	530	580	597	582	619
1	485	542	594	583	611	612
1	480	500	550	528	562	576
2	514	560	565	524	552	597
2	440	480	536	484	567	569
2	495	570	569	585	576	677
2	520	590	610	637	671	702
2	503	555	591	605	649	675
3	496	560	622	622	632	670
3	498	540	589	557	568	609
3	478	510	568	555	576	605
3	545	565	580	601	633	649
3	472	498	540	524	532	583

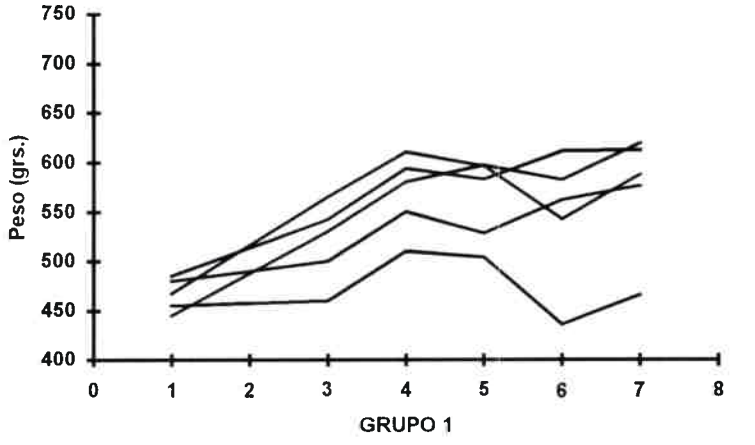
Las cuestiones de interés en este caso son:

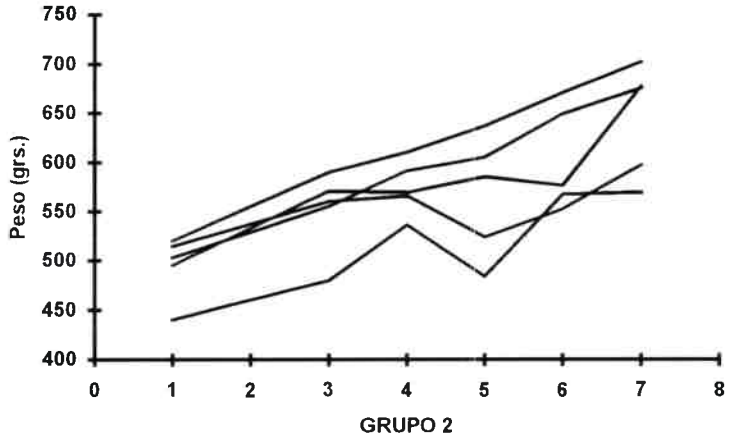
I . ¿ Tienen los perfiles de las medias de los Grupos el mismo nivel ? (es decir, ¿hay "efecto grupo" ?).

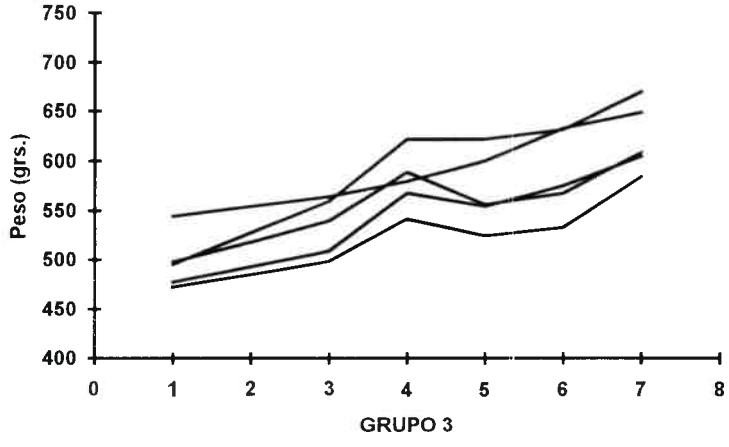
II . ¿ Son los perfiles horizontales ? (es decir, ¿hay "efecto tiempo" ?).

III . ¿ Son los perfiles paralelos ? (es decir, ¿hay "efecto interacción grupo por tiempo"?).

Inicialmente podemos examinar gráficamente el comportamiento de los datos por Grupo.

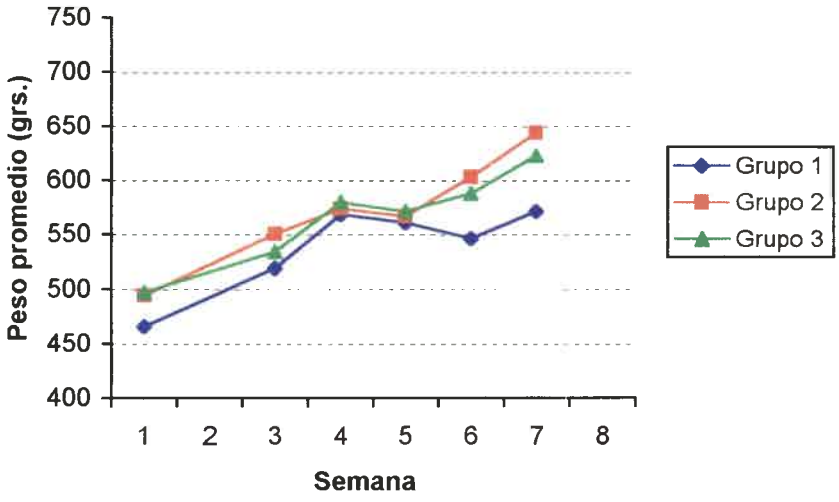






Se observa que hay variación del peso corporal a través del tiempo.

También es útil realizar un gráfico de los perfiles medios de respuesta observados para cada grupo, como el que se muestra a continuación, donde se puede ver un comportamiento muy similar entre los grupos.



Para analizar estadísticamente los datos se utilizó el paquete estadístico SAS, con el siguiente programa:

```
DATA ejem3.1;
INFILE 'a:tablaa12.prn';
INPUT grupo sem1 sem3 sem4 sem5 sem6 sem7;
PROC GLM;
  CLASS grupo;
  MODEL sem1 sem3 sem4 sem5 sem6 sem7=grupo;
  REPEATED semana 6 (1 3 4 5 6 7) polynomial / printe printm summary nou;
RUN;
```

Con este programa se indica al paquete estadístico que lea un archivo de datos (comando INFILE) y las denominaciones de las variables (comando INPUT). El PROC GLM usado realiza el análisis con las siguientes especificaciones: variable de clasificación (comando CLASS), modelo en el que hay 6 variables dependientes de la variable de clasificación (comando MODEL) y finalmente el comando REPEATED (que es específico para analizar medidas repetidas) con el que se indican el factor repetido con sus niveles, la transformación a realizar para generar contrastes ortogonales (polynomial), y que deben mostrarse la matriz con los coeficientes de correlación parcial para cada combinación de factores dentro de sujetos (printe), la matriz de transformación M que define los contrastes (printm), el análisis de la varianza para cada variable definida por las filas de M (summary), y se pide específicamente no mostrar el análisis univariado (nou).

La salida del programa es la siguiente:

1) Información general de los datos a analizar: 15 en total divididos en 3 grupos (1,2 y 3).

```
General Linear Models Procedure
    Class Level Information
   Class     Levels     Values
   GRUPO          3     1 2 3
Number of observations in data set = 15
```

2) Un análisis para cada tiempo (Semana) por separado, el cual tiene problemas al ser interpretado por la no independencia de los datos.

```
Dependent Variable: SEM1
                              Sum of           Mean
Source                  DF    Squares         Square      F Value    Pr > F
Model                    2  2969.200000  1484.600000         2.10    0.1651
Error                   12  8481.200000   706.766667
Corrected Total         14 11450.400000
                 R-Square          C.V.       Root MSE          SEM1 Mean
                 0.259310      5.467932       26.58508         486.200000

Dependent Variable: SEM3
                              Sum of           Mean
Source                  DF    Squares         Square      F Value    Pr > F
Model                    2  2497.600000  1248.800000         0.87    0.4427
Error                   12 17170.400000  1430.866667
Corrected Total         14 19668.000000
                 R-Square          C.V.       Root MSE          SEM3 Mean
                 0.126988      7.070430       37.82680         535.000000

Dependent Variable: SEM4
                              Sum of           Mean
Source                  DF    Squares         Square      F Value    Pr > F
Model                    2   302.5333333  151.2666667         0.14    0.8710
Error                   12 12992.4000000 1082.7000000
Corrected Total         14 13294.9333333
                 R-Square          C.V.       Root MSE          SEM4 Mean
                 0.022756      5.729813       32.90441         574.266667

Dependent Variable: SEM5
                              Sum of           Mean
Source                  DF    Squares         Square      F Value    Pr > F
Model                    2   260.4000000  130.2000000         0.05    0.9476
Error                   12 28906.0000000 2408.8333333
Corrected Total         14 29166.4000000
                 R-Square          C.V.       Root MSE          SEM5 Mean
                 0.008928      8.659116       49.07987         566.800000

Dependent Variable: SEM6
                              Sum of           Mean
Source                  DF    Squares         Square      F Value    Pr > F
Model                    2  8550.933333  4275.466667         1.39    0.2863
Error                   12 36898.000000  3074.833333
Corrected Total         14 45448.933333
                 R-Square          C.V.       Root MSE          SEM6 Mean
                 0.188144      9.572652       55.45118         579.266667

Dependent Variable: SEM7
                              Sum of           Mean
Source                  DF    Squares         Square      F Value    Pr > F
Model                    2 13730.13333  6865.06667         2.46    0.1276
Error                   12 33538.80000  2794.90000
Corrected Total         14 47268.93333
                 R-Square          C.V.       Root MSE          SEM7 Mean
                 0.290468      8.623338       52.86681         613.066667
```

3) El análisis de los datos usando medidas repetidas en sí: información sobre el factor repetido (Semana) y sus niveles.

Repeated Measures Level Information

Dependent Variable	SEM1	SEM3	SEM4	SEM5	SEM6	SEM7
Level of SEMANA	1	3	4	5	6	7

4) Matriz de los coeficientes de correlación parcial, en la que puede verse que hay alta correlación de cada medición con la del período anterior, y aún con la que está dos períodos antes.

Partial Correlation Coefficients from the Error SS&CP Matrix / Prob > |r|

DF = 11	SEM1	SEM3	SEM4	SEM5	SEM6	SEM7
SEM1	1.000000	0.707584	0.459151	0.543739	0.492366	0.502098
	0.0	0.0068	0.1145	0.0548	0.0874	0.0804
SEM3	0.707584	1.000000	0.889996	0.874228	0.676753	0.834899
	0.0068	0.0	0.0001	0.0001	0.0111	0.0004
SEM4	0.459151	0.889996	1.000000	0.881217	0.789575	0.847786
	0.1145	0.0001	0.0	0.0001	0.0013	0.0003
SEM5	0.543739	0.874228	0.881217	1.000000	0.803051	0.919350
	0.0548	0.0001	0.0001	0.0	0.0009	0.0001
SEM6	0.492366	0.676753	0.789575	0.803051	1.000000	0.895603
	0.0874	0.0111	0.0013	0.0009	0.0	0.0001
SEM7	0.502098	0.834899	0.847786	0.919350	0.895603	1.000000
	0.0804	0.0004	0.0003	0.0001	0.0001	0.0

5) Matriz de transformación definida por contrastes polinomiales ortogonales. La primera fila de la matriz M genera una variable derivada que representa la tendencia lineal de las curvas a través del tiempo, la segunda la tendencia cuadrática, y así siguiendo hasta el polinomio de grado 5 (pues hay 6 mediciones para cada sujeto).

SEMANA.N represents the nth degree polynomial contrast for SEMANA

M Matrix Describing Transformed Variables

	SEM1	SEM3	SEM4	SEM5	SEM6	SEM7
SEMANA.1	-.6900656	-.2760262	-.0690066	0.1380131	0.3450328	0.5520524
SEMANA.2	0.5455447	-.3273268	-.4364358	-.3273268	0.0000000	0.5455447
SEMANA.3	-.2331262	0.6061281	0.0932505	-.4196272	-.4662524	0.4196272
SEMANA.4	0.0703659	-.4817360	0.5196254	0.2760510	-.6062296	0.2219233
SEMANA.5	-.0149873	0.2248090	-.5994906	0.6744270	-.3596944	0.0749363

6) Matriz de error y los respectivos coeficientes de correlación parcial de las variables definidas por la transformación

E = Error SS&CP Matrix

SEMANA.N represents the nth degree polynomial contrast for SEMANA

	SEMANA.1	SEMANA.2	SEMANA.3	SEMANA.4	SEMANA.5
SEMANA.1	18100.87429	-981.24723	-2591.48710	-2674.32836	464.01484
SEMANA.2	-981.24723	3617.50952	-1159.36972	-1696.43738	-707.41933
SEMANA.3	-2591.48710	-1159.36972	3736.19217	1853.97859	1376.03631
SEMANA.4	-2674.32836	-1696.43738	1853.97859	2746.98421	1596.76414
SEMANA.5	464.01484	-707.41933	1376.03631	1596.76414	4351.03980

Partial Correlation Coefficients from the Error SS&CP Matrix of the Variables Defined by the Specified Transformation / Prob > |r|

DF = 11	SEMANA.1	SEMANA.2	SEMANA.3	SEMANA.4	SEMANA.5
SEMANA.1	1.000000	-0.121262	-0.315126	-0.379260	0.052286
	0.0	0.6931	0.2943	0.2012	0.8653
SEMANA.2	-0.121262	1.000000	-0.315357	-0.538152	-0.178310
	0.6931	0.0	0.2939	0.0578	0.5600
SEMANA.3	-0.315126	-0.315357	1.000000	0.578711	0.341286
	0.2943	0.2939	0.0	0.0382	0.2538
SEMANA.4	-0.379260	-0.538152	0.578711	1.000000	0.461866
	0.2012	0.0578	0.0382	0.0	0.1121
SEMANA.5	0.052286	-0.178310	0.341286	0.461866	1.000000
	0.8653	0.5600	0.2538	0.1121	0.0

7) Test de esfericidad (cuya interpretación se vé en el análisis univariado)

Test for Sphericity: Mauchly's Criterion = 0.0544835
Chisquare Approximation = 29.389556 with 14 df Prob > Chisquare = 0.0093

8) Tests multivariados para el efecto SEMANA

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no SEMANA Effect
H = Type III SS&CP Matrix for SEMANA E = Error SS&CP Matrix
S=1 M=1.5 N=3

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.03881848	39.6175	5	8	0.0001
Pillai's Trace	0.96118152	39.6175	5	8	0.0001
Hotelling-Lawley Trace	24.76092347	39.6175	5	8	0.0001
Roy's Greatest Root	24.76092347	39.6175	5	8	0.0001

9) Tests multivariados para el efecto de la interacción SEMANAxGRUPO

Manova Test Criteria and F Approximations for
the Hypothesis of no SEMANA*GRUPO Effect
H = Type III SS&CP Matrix for SEMANA*GRUPO E = Error SS&CP Matrix
S=2 M=1 N=3

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.17905151	2.1812	10	16	0.0793
Pillai's Trace	1.07058517	2.0734	10	18	0.0856
Hotelling-Lawley Trace	3.19076786	2.2335	10	14	0.0824
Roy's Greatest Root	2.66824588	4.8028	5	9	0.0205

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

10) Test para el efecto GRUPO

Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GRUPO	2	18548.067	9274.033	1.06	0.3782
Error	12	105434.200	8786.183		

11) Análisis para los coeficientes de los polinomios ortogonales.

Analysis of Variance of Contrast Variables

SEMANA.N represents the nth degree polynomial contrast for SEMANA

Contrast Variable: SEMANA.1

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MEAN	1	131764.80286	131764.80286	87.35	0.0001
GRUPO	2	2495.21333	1247.60667	0.83	0.4608
Error	12	18100.87429	1508.40619		

Contrast Variable: SEMANA.2

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MEAN	1	2011.4793651	2011.4793651	6.67	0.0240
GRUPO	2	4489.6777778	2244.8388889	7.45	0.0079
Error	12	3617.5095238	301.4591270		

Contrast Variable: SEMANA.3

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MEAN	1	2862.1936232	2862.1936232	9.19	0.0104
GRUPO	2	694.1098551	347.0549275	1.11	0.3597
Error	12	3736.1921739	311.3493478		

Contrast Variable: SEMANA.4

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MEAN	1	3954.8810579	3954.8810579	17.28	0.0013
GRUPO	2	1878.3636040	939.1818020	4.10	0.0439
Error	12	2746.9842142	228.9153512		

Contrast Variable: SEMANA.5

Source	DF	Type III SS	Mean Square	F Value	Pr > F
MEAN	1	1961.1430967	1961.1430967	5.41	0.0384
GRUPO	2	205.3687631	102.6843816	0.28	0.7583
Error	12	4351.0398023	362.5866502		

Se analiza en primer lugar el efecto de la interacción, para lo cual en 9), aunque los valores de los estadísticos y de los p del test difieren (por haber más de dos grupos), a un nivel del 1% no resultan significativos; por esto podemos asumir que no hay interacción SEMANAxGRUPO y se pueden observar los efectos principales. En cuanto a las preguntas enunciadas al principio, estamos respondiendo a III, y concluimos que los perfiles son paralelos.

Para la pregunta I recurrimos a 10), en donde se presenta un valor p del test de 0.3782 que indica que puede asumirse que no hay efecto GRUPO, es decir que no hay diferencia estadísticamente significativa entre los pesos medios con las tres dietas.

En cuanto al efecto SEMANA (pregunta II), observamos 8), en donde se ve que todos los tests proveen el mismo valor F y por ende el mismo p del test. En estos tests se prueban simultáneamente las variables derivadas lineal, cuadrática, cúbica, de orden cuatro y cinco, para ver si sus medias son significativamente diferentes de cero, y si alguna lo es los resultados no son constantes a través del tiempo. En este caso todos los tests muestran que el efecto SEMANA es altamente significativo.

Para ver la significación de los polinomios separádamente recurrimos a 11), y observamos en cada análisis la fila "MEAN". En ellas los valores del p del test llevan a concluir que resultan significativas todas las variables derivadas, por lo que no alcanza sólo con una recta, ni con una parábola, etc., sino que se necesita un polinomio de grado 5 para explicar los cambios con el tiempo. Si queremos conocer los estimadores de los efectos de cada grado debemos multiplicar los valores de la matriz M dada en 5) por las medias de cada grupo en cada semana, y luego promediar los valores para los tres grupos.

CAPITULO 3
Análisis de la Varianza Univariado

El enfoque univariado de datos provenientes de medidas repetidas es muy habitual en varias disciplinas en las que hay historia en el uso de los métodos de análisis de la varianza. A simple vista parece que los n individuos medidos en p ocasiones forman un diseño en Bloques aleatorizados o en Parcelas divididas, por lo que debería llevarse a cabo un análisis de la varianza de dos factores.

En esos análisis las np observaciones son consideradas como componentes de un solo vector de datos npx1: el primer conjunto de p valores son las p observaciones del primer individuo, el segundo conjunto corresponde al segundo individuo, y así sucesivamente.

Ejemplo 3.1: Consideraremos los datos del Ejemplo 2.8, y los analizaremos desde el punto de vista univariado. Hay dos grupos de tres y dos individuos, respectivamente, con k perfiles de respuesta esperados definidos en función a los parámetros μ_{kj} como $(\mu_{k1}+\mu_{k2} \quad \mu_{k1}+\mu_{k3} \quad \mu_{k1}-\mu_{k2}-\mu_{k3})$. Entonces

$$\mathrm{E}\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{41} \\ y_{42} \\ y_{43} \\ y_{51} \\ y_{52} \\ y_{53} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix}$$

El modelo apropiado para describir estos datos parece ser

$$y_{ij} = \alpha_i + \beta_j + u_{ij} \qquad \textbf{(3.1)}$$

donde y_{ij} es la observación del i-ésimo sujeto en el j-ésimo tiempo

α_i es el efecto sujeto

β_j es el efecto tiempo

u_{ij} es el error aleatorio

Hay dos problemas para aplicar directamente este modelo a los datos.

El primer problema es que en general en un Diseño en Bloques o en Parcelas Divididas, los α_i corresponden a categorías prefijadas y son por lo tanto efectos fijos, lo que implica que las np observaciones son independientes y por ende la matriz de covarianza asociada a ellas es diagonal. En medidas repetidas en cambio, los α_i son aleatoriamente seleccionados de una distribución de efectos de los individuos, siendo por ende efectos aleatorios, por lo que la matriz antes mencionada resulta ser diagonal por bloques.

Más formalmente

$$y_{ij} = \alpha + a_i + \beta_j + e_{ij} \qquad (i = 1, ..., n;\ j = 1, ..., p) \qquad \textbf{(3.2)}$$

donde se ha cambiado α_i por $\alpha + a_i$ (con a_i un efecto aleatorio y α una constante para todo i, j), y u_{ij} por e_{ij} para indicar que hay más de una fuente de variación.

Las diferencias entre individuos se manifiestan en la variación aleatoria entre individuos modelada por a_i.

Sea a_i tal que $\mathrm{E}(a_i) = 0$ y $\mathrm{Var}(a_i) = \sigma^2_{a_i}$, e_{ij} tal que $\mathrm{E}(e_{ij}) = 0$ y $\mathrm{Var}(e_{ij}) = \sigma^2$ y ambos con distribuciones independientes, entonces

$$\begin{aligned}\mathrm{cov}(y_{ij}, y_{ij'}) &= \mathrm{E}\left[(a_i + \beta_j + e_{ij})(a_i + \beta_{j'} + e_{ij'})\right] \\ &\quad - \mathrm{E}(a_i + \beta_j + e_{ij})\mathrm{E}(a_i + \beta_{j'} + e_{ij'}) \\ &= \sigma^2_a\end{aligned}$$

$$\text{cov}(y_{ij}, y_{ij}) = \sigma_a^2 + \sigma^2$$

$$\text{cov}(y_{ij}, y_{i'j}) = 0$$

La matriz de covarianza $\boldsymbol{\Sigma}$ de las np observaciones consiste en una serie de npxp matrices diagonales por bloque, cada una de las cuales tiene como elementos de la diagonal a $\sigma_a^2 + \sigma^2$ y fuera de la ella σ_a^2.

Definición 3.1: Cada una de las submatrices que componen la matriz de covarianza $\boldsymbol{\Sigma}$ con las especificaciones anteriores, se dice que es ***simétrica compuesta***. Podemos escribir esto como

$$\boldsymbol{\Sigma}_p = \sigma_a^2 \mathbf{1}\mathbf{1}^T + \sigma^2 \boldsymbol{I}$$

donde $\mathbf{1}$ es un vector de p "unos".

Esto significa que para cada sujeto hay homogeneidad de varianzas (cada una vale $\sigma_a^2 + \sigma^2$) , mientras que las covarianzas también son iguales entre sí y valen σ_a^2.

Por lo tanto una matriz de covarianza simétrica compuesta para cuatro variables tiene la estructura

$$\begin{pmatrix} \sigma_a^2 + \sigma^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 + \sigma^2 & \sigma_a^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma^2 & \sigma_a^2 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma^2 \end{pmatrix}$$

Nota: Aunque $\boldsymbol{\Sigma}$ no es diagonal cuando los individuos constituyen un efecto aleatorio, los tests F usuales siguen siendo válidos, y lo son en realidad bajo la siguiente condición más general.

Definición 3.2: Se dice que se verifica la ***esfericidad*** o ***circularidad*** cuando $\text{Var}(y_{ij} - y_{ij'})$ es constante para todo $j \neq j'$, lo cual dice que la matriz de covarianza de cualquier conjunto de $p-1$ contrastes ortonormales en cada ocasión debe ser proporcional a la matriz identidad.

Que la simetría compuesta es un caso particular de esfericidad puede verse claramente de la siguiente manera:
como

$$\text{Var}(y_{ij} - y_{ij'}) = \text{Var}(y_{ij}) + \text{Var}(y_{ij'}) - 2\text{Cov}(y_{ij}, y_{ij'}).$$

cuando la simetría compuesta se cumple resulta que

$$\text{Var}(y_{ij} - y_{ij'}) = (\sigma_a^2 + \sigma^2) + (\sigma_a^2 + \sigma^2) - 2\sigma_a^2 = 2\sigma^2$$

que es constante como se requiere para que valga la esfericidad.

Cada elemento de la diagonal principal de la matriz de covarianza de los contrastes ortonormales estima el cuadrado medio que es el denominador en el test F de ese contraste. Por lo tanto, si se cumple la esfericidad, cada uno estima lo mismo, y en ese caso puede obtenerse un mejor estimador promediando sus valores (que es lo que el test F usual hace). Esto explica porqué el test F es llamado muchas veces el "test-F promedio" en la salida de algunos paquetes estadísticos para medidas repetidas. Si la esfericidad no se cumple, se estarían promediando diferentes cosas, y por ende los tests resultantes de contrastes individuales podrían tener denominadores muy grandes o muy pequeños, lo que lleva a que sean sesgados. Si sólo ciertos contrastes son de interés se puede restringir la atención a la esfericidad de éstos, lo que implica reducir los grados de libertad en la suma de cuadrados del denominador, resultando menos potente pero a la vez con condiciones menos restrictivas.

El segundo problema al que hacíamos referencia al comenzar a tratar las mediciones repetidas usando un análisis univariado, concierne al orden natural de las observaciones. A diferencia del Diseño en Bloques o en Parcelas Divididas, en este caso no pueden aleatorizarse las ocasiones (la primera medición es la primera y no puede ser ninguna otra). Similarmente la primera y segunda medición estarán más cercanas que, por ejemplo, la primera y la tercera; en consecuencia puede esperarse mayor correlación entre las mediciones cercanas que entre las que están alejadas temporalmente. La estructura de simetría compuesta puede ser reemplazada por otra, por ejemplo una matriz en la que las correlaciones decrecen a medida que aumenta la

distancia a la diagonal principal, lo que implica que el test F usual ya no es válido.

El modelo (3.2) expresa cada observación como una combinación lineal de efectos aleatorios (los a_i y e_{ij}) y efectos fijos (los β_j). Los efectos aleatorios en sí tienen estructuras de covarianza particularmente simples. Puede expresarse dicho modelo alternativamente como $y_{ij} = \beta_j + u_{ij}$, donde β_j sigue siendo un efecto fijo y el efecto simple aleatorio $(u_{i1} \quad \ldots \quad u_{ip})$ tiene asociada una matriz de covarianza simétrica compuesta. Esto lleva a una natural generalización que, escrita en términos matriciales es:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} \qquad \textbf{(3.3)}$$

donde $\boldsymbol{X}$ es la matriz $(np\text{x}q)$ del diseño

$\boldsymbol{\beta}$ es un vector $(q\text{x}1)$ de parámetros fijos

$\boldsymbol{u}$ es un vector $(np\text{x}1)$ de términos aleatorios con media cero y matriz de covarianza $\boldsymbol{\Sigma}$ que es "diagonal por bloques" con bloques simétricos compuestos.

Por esto resulta que

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{y}$$

Ya no requerimos que $\boldsymbol{\Sigma}_p$ sea simétrica compuesta.

Dado que este análisis univariado es válido cuando la esfericidad (o la simetría compuesta) se cumple, llegamos naturalmente a dos preguntas:

1) ¿ cómo probar si una matriz satisface la esfericidad ?

2) ¿ qué hacer si no la satisface (y aún quisiéramos usar un análisis univariado)?

Tests y medidas para la esfericidad

Existen tests para probar la simetría compuesta, pero en general los programas estadísticos verifican el supuesto más general de esfericidad.

Muchos programas, como los que en esta monografía se usan, muestran el resultado de **test de Mauchly** de esfericidad, cuyas hipótesis son:

H_0: Se cumple la esfericidad $\qquad$ H_1 : No se cumple la esfericidad

y cuyo estadístico es:

$$W = \frac{|\mathbf{cs}_p\mathbf{c}^T|}{|\text{tr}(\mathbf{cs}_p\mathbf{c}^T)/(p-1)|^{p-1}} \qquad \textbf{(3.4)}$$

donde $\mathbf{S}_p$ es la matriz de covarianza muestral dentro de sujetos combinada y $\mathbf{C}$ es una matriz de los $p-1$ contrastes ortonormales.

Este test está basado en

$$\{\Sigma_g(n_g-1)-(p-2)(p+1)/2\}\log(W)$$

con n_g : número de datos del grupo g, que está asintóticamente distribuída como una χ^2_ν con $\nu=(p-2)(p+1)/2$ cuando la hipótesis nula es cierta.

Si la esfericidad no se cumple entonces el test F usual rechaza muchas hipótesis nulas verdaderas. Por esto se ha definido una medida de la desviación a la esfericidad denominada ϵ, que sirve también para ajustar los tests F usuales.

Sea $\mathbf{\Sigma}_c=\mathbf{C}\mathbf{\Sigma}_p\mathbf{C}^T$ la matriz de covarianza para un conjunto de $p-1$ contrastes ortonormales. Entonces:

$$\epsilon=\frac{\{\text{tr}(\Sigma_c)\}^2}{(p-1)\text{tr}(\Sigma_c^2)}=\frac{(\Sigma\theta_j)^2}{(p-1)\Sigma\theta_j^2} \qquad \textbf{(3.5)}$$

donde los θ_j son los $p-1$ autovalores de $\mathbf{\Sigma}_c$.

Según Hand y Crowder (1996) es de destacar la analogía entre ϵ y la varianza: ésta es la diferencia entre el promedio de los cuadrados $\Sigma x_i^2/n$ y el cuadrado de los promedios $(\Sigma x_i/n)^2$ mientras que ϵ es la razón entre el cuadrado del autovalor promedio $(\Sigma\theta_j/(p-1))^2$ y el promedio de los cuadrados de los autovalores $\Sigma\theta_j^2/(p-1)$.

Cuando la esfericidad se cumple $\epsilon = 1$, de otra manera $\epsilon < 1$. Esto se debe a que los θ_j son todos iguales sólo si la esfericidad es válida (de la misma manera que la varianza es cero sólo sí los valores son todos iguales). Cuando el supuesto no se cumple se pueden realizar los ajustes presentados a continuación.

Ajustes para la no esfericidad

En el caso de G grupos, con un total de n individuos cada uno medido en p tiempos, el análisis de la varianza usual para probar el efecto "tiempo" asume que el estadístico construído como la razón de los cuadrados medios tiene distribución F con $p-1$ y $(p-1)(n-G)$ grados de libertad, lo cual es adecuado cuando se cumple la esfericidad. Si este supuesto no es válido, el estadístico tiene también distribución F pero con grados de libertad ajustados multiplicando por ϵ los grados de libertad del test usual. Esto puede dar como resultado grados de libertad fraccionarios, de tal manera que al buscar en tablas "manualmente" deben hacerse interpolaciones.

Este ajuste sólo se realiza para probar el efecto tiempo y de la interacción (el test para el efecto grupo permanece sin cambios).

Sin embargo el valor de ϵ es desconocido, y al estimarlo se introducen errores que afectan las distribuciones.

Un estimador obvio se logra sustituyendo $\boldsymbol{\Sigma}$ en la ecuación (3.5) por su estimador muestral. Éste es el **estimador de Greenhouse-Geisser** que muchos programas estadísticos informan. En realidad Greenhouse y Geisser (1959) sugirieron la siguiente estrategia para solucionar el problema de la variación intrínseca en su estimador:

1 . Realizar el test F con grados de libertad no ajustados. Si dá un resultado no significativo entonces terminar (pues el ajuste no daría un resultado significativo). De otra manera seguir con el paso 2.

2 . Si el estadístico no ajustado F resulta significativo, realizar un test F ajustado conservativo, ajustando los grados de libertad multiplicando por

$1/(p-1)$ que es el mínimo valor de ϵ. Si este resultado es significativo frenar (el test F ajustado por ϵ da un resultado entre el obtenido en el paso 1 y el del paso 2, por lo que si ambos fueron significativos aquel también lo sería). Si éste no es el caso seguir con el paso 3.

3 . Estimar ϵ usando la matriz de covarianza muestral y realizar el test con el ajuste por $\widehat{\epsilon}$.

Trabajos posteriores han sugerido que para $\epsilon > 0.75$ y $n < 2p$ el estimador simple de $\epsilon, \widehat{\epsilon}$ puede estar seriamente sesgado de manera tal que tiende a sobrecorregir los grados de libertad y producir un test conservativo.

Por esto ***Huynh y Feldt (1976)*** definieron un estimador $\widetilde{\epsilon}$ para ϵ que es menos sesgado que $\widehat{\epsilon}$ cuando $\epsilon > 0.75$. Estos valores de ϵ son comunes en los estudios en áreas como psicología y educación. La expresión para $\widetilde{\epsilon}$ está dada por :

$$\widetilde{\epsilon} = \min(1, \frac{a}{b}) \qquad a = n(p-1)\widehat{\epsilon} - 2$$

$$b = (p-1)\{n - G - (p-1)\widehat{\epsilon}\}$$

con $\widehat{\epsilon}$ el estimador simple de ϵ usando la matriz de covarianza muestral. Se sugiere en realidad usar $\widetilde{\epsilon}$ como estimador para ϵ sólo cuando se cree que $\epsilon > 0.75$.

Aunque hay una inclinación a usar el test de Mauchly para probar esfericidad, y si ésta no se cumple realizar los ajustes correspondientes, parece ser que éste es un test débil para detectar pequeños apartamientos de la esfericidad, lo cual implicaría sesgos sustanciales en los test F usuales. Esto sugiere que sería necesario realizar siempre los tests ajustados, a menos que se tengan razones teóricas para suponer que la esfericidad se mantiene.

Enfoque univariado del Problema 1

Usando los datos del peso corporal de cerdos "guinea" ya analizados de "manera multivariada" en el Problema 1, realizaremos el análisis univariado de los mismos para responder a las cuestiones allá enunciadas.

El programa en SAS para el análisis de los datos es muy similar al presentado anteriormente, salvo que se cambia la opción "nou" del comando REPEATED por "nom", lo cual indica al programa que no se muestre el análisis multivariado.

En la salida aparecen los items denotados de 1) a 7) en el enfoque multivariado, y a continuación el análisis univariado siguiente.

8) Test para el efecto GRUPO (que coincide con el dado en el análisis multivariado).

Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GRUPO	2	18548.067	9274.033	1.06	0.3782
Error	12	105434.200	8786.183		

9) Tests univariados usuales para los efectos SEMANA y de la interacción SEMANAxGRUPO, con el agregado de los valores p del test para los ajustes por no esfericidad y los valores de los estimadores de ϵ.

Univariate Tests of Hypotheses for Within Subject Effects

Source: SEMANA

DF	Type III SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
5	142554.5000000	28510.9000000	52.55	0.0001	0.0001	0.0001

Source: SEMANA*GRUPO

DF	ype III SS	Mean Square	F Value	Pr > F	Adj Pr > F G - G	Adj Pr > F H - F
10	9762.7333333	976.2733333	1.80	0.0801	0.1457	0.1103

Source: Error(SEMANA)

DF	Type III SS	Mean Square
60	32552.6000000	542.5433333

Greenhouse-Geisser Epsilon = 0.4856
Huynh-Feldt Epsilon = 0.7191

A continación puede verse el punto 11) de la salida del enfoque multivariado.

En cuanto a la interpretación de la salida, Hand y Crowder (1996) sugieren observar las sumas de cuadrados de los componentes de los polinomios ortogonales y la matriz de correlación de los mismos. Esto puede verse en 11) y en la segunda parte de 6) de la salida del enfoque multivariado, respectivamente. En 11) se ve que las sumas de cuadrados son muy diferentes entre sí, lo que sugiere que la matriz de covarianza es muy distinta de la identidad. También se confirma esta conclusión con la matriz de correlación que presenta coeficientes sustancialmente diferentes.

Por otra parte el item 7) es el test de esfericidad el cual debe satisfacerse para los componentes de los polinomios ortonormales, para que el análisis univariado usual sea válido. La salida de este test es:

```
          Test for Sphericity: Mauchly's Criterion = 0.0544835
Chisquare Approximation = 29.389556 with 14 df    Prob > Chisquare = 0.0093
```

La hipótesis nula plantea que hay esfericidad, el estadístico es $W = 0.0544835$ y el valor p del test es 0.0093, lo que lleva a rechazar el cumplimiento del supuesto de esfericidad (confirmando lo observado en las matrices).

Por esto en vez de utilizar el análisis de la varianza usual, debemos realizar los ajustes de Greenhouse-Geisser y Huynh-Feldt.

Observamos en la salida 9) (del presente enfoque) en primer lugar el efecto interacción SEMANAxGRUPO, para el cual los valores p del test ajustados por Greenhouse-Geisser y Huynh-Feldt son respectivamente 0.1457 y 0.1103, los cuales indican que la interacción no es significativa (el valor p del test no ajustado 0.0801 dejaba ciertas dudas sobre la decisión a tomar).

En cuanto al efecto GRUPO puede verse nuevamente 8) de la salida del enfoque multivariado, dado que para ese efecto no son necesarios los ajustes. Concluimos que (como $p > 0.05$) los Grupos no presentan diferencias estadísticamente significativas en cuanto a los perfiles de respuesta (conclusión idéntica a la obtenida en el análisis multivariado).

A continuación podemos ver el efecto SEMANA, para el cual el valor p del test resulta ser el mismo para el test no ajustado y los dos ajustes (0.0001) que indica que el efecto SEMANA es altamente significativo. Como en el caso multivariado, cabe aclarar que con este test estamos probando simultáneamente la significación de las variables derivadas lineal, cuadrática, cúbica, de orden cuatro y cinco, para ver si sus medias son significativamente diferentes de cero y, si alguna lo es, los resultados no son constantes a través del tiempo. En este caso todos los tests muestran que el efecto SEMANA es altamente significativo. Entonces para ver la significación de los polinomios separádamente recurrimos a 11) de la salida del enfoque multivariado del Problema 1 y las conclusiones son las obtenidas en ese caso.

Finalmente un comentario acerca de los estimadores de ϵ: los valores son 0.4856 y 0.7119 para Greenhouse-Geisser y Huynh-Feldt, respectivamente, los cuales al estar alejados de uno, llevan a pensar en el no cumplimiento de la esfericidad (que se confirma con el test).

Enfoques univariado y multivariado del Problema 1 con STATISTICA

A continuación se presenta un análisis de los datos del Problema 1 usando el paquete STATISTICA, con el fin de mostrar que ambos paquetes (SAS y STATISTICA) brindan aproximadamente la misma información, con distinta presentación.

Se ingresaron los datos en el mismo formato anterior pero con extensión STA. Para realizar el análisis, se ingresó en primer lugar a la opción *ANOVA/MANOVA* (que es el "entorno" en el que se realiza todo el trabajo) del menú *STATISTICA Module Switcher*. Se pueden realizar en primer lugar gráficos similares a los presentados anteriormente para este problema, ingresando a la opción *Graphs* y a continuación a la opción *Stats 2D Graphs*, y dentro de ésta en *Line Plots (Case Profiles)*. Con la finalidad de evitar redundancia no se muestran en este caso, aunque sí se exhibirán para los datos del Problema 2 (en aquél se indicará mejor cómo realizar los gráficos en STATISTICA).

Para indicar que se trabajará con medidas repetidas en la opción *Analysis*, se ingresó en *Startup Panel*. En esa ventana se determinan la variable independiente "Grupo" y las dependientes "S1, S3, S4, S5, S6 y S7" (en la opción *Variables*). En la opción *Codes for between-groups factors* con *OK* se indica tomar como niveles de la variable Grupo los indicados (1, 2 y 3). Finalmente en *Repeated-measures (within SS) design* se indica que se tienen 6 niveles para el factor repetido al que se denomina (en este caso) Semana.

En la ventana *ANOVA Results:*
Eligiendo la opción *All-effects* se puede ver:

1)

Summary of all Effects; design: (dieta2.sta)
1-GRUPO, 2-SEMANA

Effect	df Effect	MS Effect	df Error	MS Error	F	p-level
1	2	9274.03	12	8786.184	1.05552	.378209
2	5	28510.90	60	542.543	52.55046	.000000
12	10	976.27	60	542.543	1.79944	.080144

que es el análisis univariado para Grupo, Semana y Grupo x Semana, respectivamente.

En esta salida las informaciones que figuran en el renglón correspondiente al efecto Grupo coinciden con las brindadas por SAS en 10) del enfoque multivariado del Problema 1 y en 8) del enfoque univariado (ya que Grupo tiene sólo dos niveles). El segundo renglón del efecto Semana se corresponde con parte del punto 9) del Enfoque univariado del Problema 1, específicamente con el análisis no ajustado por esfericidad mostrado en la primera tabla ANOVA. Análogamente, para el renglón correpondiente a la interacción debemos remitirnos a la segunda tabla ANOVA del mismo punto, y las conclusiones son por lo tanto las detalladas anteriormente.

Si se opta por *Means-effects* se presenta la tabla de la salida 1) anterior y se puede elegir cualquiera de los tres renglones.

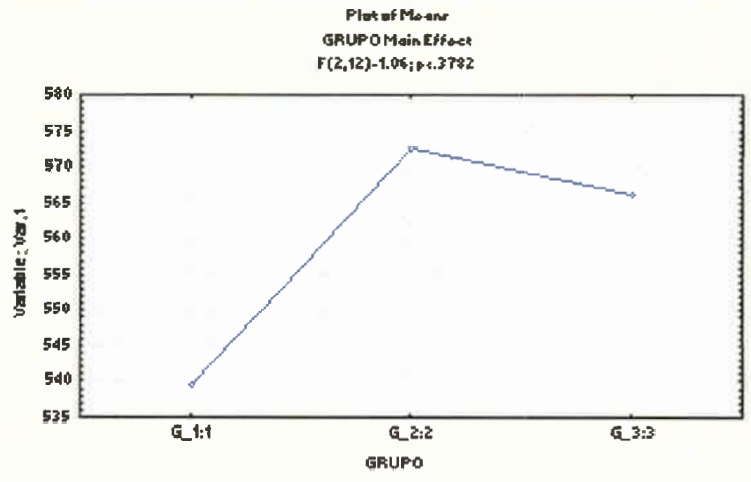
Eligiendo el renglón "Grupo" y optando por *Scrollsheet* aparece

2)a)i)

```
Means (dieta2.sta)
F(2,12)=1.06; p<.3782
    Depend.
    Var.1
1      ....      539.1334
2      ....      572.2667
3      ....      565.9000
```

que muestra la media de la variable peso en cada grupo. Si se elige *Graphs* aparece

2)a)ii)



en el que puede verse cláramente que en promedio el Grupo 2 parece ser el mejor. La marcada diferencia gráfica (no detectada por el test para Grupo) se debe a la escala "exageradamente grande" usada.

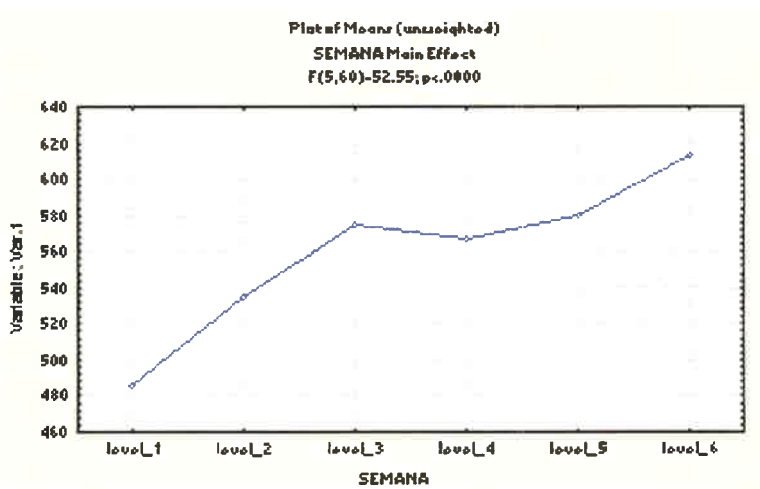
Si se elige el renglón "Semana" y se opta por *Scrollsheet* aparece

2)b)i)

```
Means (unweighted) (dieta2.sta)
Rao R (5,8)=39.62; p<.0000
    Depend.
    Var.1
....          1  486.2000
....          2  535.0000
....          3  574.2667
....          4  566.8000
....          5  579.2667
....          6  613.0667
```

donde se muestran las medias por semana. Eligiendo *Graphs*

2)b)ii)



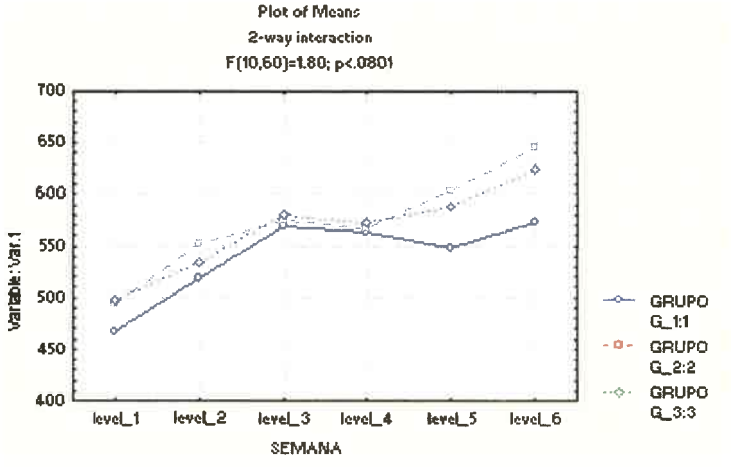
Si se elige el renglón "Grupo x Semana" y se opta por *Scrollsheet*, se vé

2)c)i)

```
Means (dieta2.sta)
Rao R (10,16)=2.18; p<.0793
    Depend.
    Var.1
1         1              466.4000
1         2              519.4000
1         3              568.8000
1         4              561.6000
1         5              546.6000
1         6              572.0000
2         1              494.4000
2         2              551.0000
2         3              574.2000
2         4              567.0000
2         5              603.0000
2         6              644.0000
3         1              497.8000
3         2              534.6000
3         3              579.8000
3         4              571.8000
3         5              588.2000
3         6              623.2000
```

y si se elige *Graphs* indicando colocar en el eje x al factor Semana, aparece el siguiente gráfico que ilustra la gran similitud entre las curvas (no cruce de líneas y paralelismo de segmentos), lo que ya fue probado con un valor $p > 0.05$ para la interacción.

2)c)ii)



Otra manera de ver las tablas ANOVA para los efectos específicos, es con la opción *Specific effect/Means/Graphs*; eligiendo el efecto Grupo aparece la siguiente tabla que es una "ampliación" del primer renglón de 1).

3)a)

MAIN EFFECT: GRUPO (dieta2.sta)
1-GRUPO, 2-SEMANA

	Sum of Squares	df	Mean Square	F	p-level
Effect	18548.1	2	9274.033	1.055525	.378209
Error	105434.2	12	8786.184		

Si en cambio se elige el efecto Semana aparece

3)b)

MAIN EFFECT: SEMANA (dieta2.sta)
1-GRUPO, 2-SEMANA

	Value	p-level
Wilks' Lambda	.03882	
Rao R Form 2 (5, 8)	39.61748	.000020
Pillai-Bartlett Trace	.96118	
V (5,8)	39.61748	.000020

que contiene parte de la información dada por SAS que se mostró en 8) del enfoque multivariado de este problema (sólo dos de los cuatro tests presentados en ese caso). En realidad los p del test tienen valores a primera vista diferentes (por la cantidad de decimales usados) aunque llevan a las mismas conclusiones.

De la misma manera eligiendo la interacción se puede ver

3)c)

INTERACTION: 1 x 2 (dieta2.sta)
1-GRUPO, 2-SEMANA

	Value	p-level
Wilks' Lambda	.179052	
Rao R Form 3 (10, 16)	2.181212	.079316
Pillai-Bartlett Trace	1.070585	
V (10,18)	2.073405	.085566

que es similar a 9) del enfoque multivariado, con los mismos valores para los p del test.

Si en *Output Options* se seleccionan los ajustes de Greenhouse/Geisser & Huynh/Feldt y el test de Mauchly, y luego se ingresa en *Specific effect / Means / Graphs*, al elegir "Grupo" aparece nuevamente 3)a), pero al elegir "Semana" aparecen (entre otras cosas):

4)a)

Mauchly's Sphericity Test (dieta2.sta)
MAIN EFFECT: SEMANA

	W	Chi-Sqr.	df	p
Sphericity Test	.054484	29.38956	14	.009275

que coincide con 7) del enfoque multivariado (aunque en aquél caso no se analiza).

También aparecen los estimadores de Epsilon, que están mostrados al final de 9) del enfoque univariado.

4)b)

Greenhouse/Geisser & Huynh/Feldt Epsilon (dieta2.sta)
MAIN EFFECT: SEMANA

	Epsilon
Greenhouse-Geisser Epsilon	.485573
Huynh-Feldt Epsilon	.719129
Lower-bound Epsilon	.200000

Nuevamente puede verse una tabla ampliada de la 1) para Semana.

4)c)

MAIN EFFECT: SEMANA (dieta2.sta)
1-GRUPO, 2-SEMANA

	Sum of Squares	df	Mean Square	F	p-level
Effect	142554.5	5	28510.90	52.55046	.000000
Error	32552.6	60	542.54		

y también otra tabla ANOVA con los ajustes que es la primera tabla de 9) del enfoque univariado, con algunos pequeños cambios en los valores del p del test.

4)d)

Univariate Test with Adjusted Degrees of Freedom (dieta2.sta)
F = 52.55046
MAIN EFFECT: SEMANA

	Unadjstd	Greenhs. Geisser	Huynh Feldt	Lower Bound
Epsilon		.48557	.71913	.20000
df 1	5.00000	2.42787	3.59564	1.00000
df 2	60.00000	29.13439	43.14772	12.00000
p-level	.00000	.00000	.00000	.00001

Si en *Specific effect/Means/Graphs*, se opta por la interacción, vuelven a aparecer 4)a) y 4)b) idénticos, y los análisis análogos a 4)c) y 4)d) son los siguientes:

5)

INTERACTION: 1 x 2 (dieta2.sta)
1-GRUPO, 2-SEMANA

	Sum of Squares	df	Mean Square	F	p-level
Effect	9762.73	10	976.2733	1.799438	.080144
Error	32552.60	60	542.5433		

Univariate Test with Adjusted Degrees of Freedom (dieta2.sta)
F = 1.799438
INTERACTION: 1 x 2

	Unadjstd	Greenhs. Geisser	Huynh Feldt	Lower Bound
Epsilon		.48557	.71913	.20000
df 1	10.00000	4.85573	7.19129	2.00000
df 2	60.00000	29.13439	43.14772	12.00000
p-level	.08014	.14428	.11212	.20727

Si se opta por *Descriptive Statistics & Graphs*, y en *Descriptive Statistics* se elige *Pooled within-groups covar's/corr's* aparecen las siguientes matrices

6)a)

Pooled Within-Groups Covariances (dieta2.sta)

	S1	S3	S4	S5	S6	S7
S1	706.7667	711.567	401.650	709.467	725.833	705.683
S3	711.5667	1430.867	1107.750	1623.033	1419.517	1669.617
S4	401.6500	1107.750	1082.700	1423.117	1440.650	1474.767
S5	709.4667	1623.033	1423.117	2408.833	2185.533	2385.433
S6	725.8333	1419.517	1440.650	2185.533	3074.833	2625.483
S7	705.6833	1669.617	1474.767	2385.433	2625.483	2794.900

6)b)

Pooled Within-Groups Correlations (dieta2.sta)

	S1	S3	S4	S5	S6	S7
S1	1.000000	.707584	.459151	.543739	.492366	.502098
S3	.707584	1.000000	.889996	.874228	.676753	.834899
S4	.459151	.889996	1.000000	.881217	.789575	.847786
S5	.543739	.874228	.881217	1.000000	.803051	.919350
S6	.492366	.676753	.789575	.803051	1.000000	.895603
S7	.502098	.834899	.847786	.919350	.895603	1.00000

donde 6)b) contiene parte de la información dada en 4) del enfoque multivariado (sólo aparecen los coeficientes de correlación parcial, no los valores p del test de los mismos).

En la opción *Tests of ANOVA Assumptions* eligiendo *Homogeneity of Variances/Covariances* aparece

7)a)

Tests of Homogeneity of Variances (dieta2.sta)

	Hartley F-max	Cochran C	Bartlett Chi-sqr	df	p
S1	3.639385	.480262	1.476551	2	.477945
S3	1.981260	.408843	.474374	2	.788845
S4	1.999745	.482497	.508639	2	.775446
S5	2.501461	.532969	.900610	2	.637438
S6	2.341535	.484883	.648711	2	.722996
S7	3.054268	.455771	1.155926	2	.561046

7)b)

Levene's Test for Homogeneity of Variances (dieta2.sta)
(ANOVA on absolute within-cell deviation scores)
Degrees of freedom for all F's: 2,12

	MS Effect	MS Error	F	p-level
S1	97.1624	287.9694	.337405	.720176
S3	62.2190	431.1414	.144312	.867097
S4	173.6001	319.2947	.543699	.594248
S5	469.5793	375.6463	1.250057	.321262
S6	180.5537	800.3670	.225589	.801355
S7	508.2559	702.1827	.723823	.504905

que muestran la homogeneidad de varianza en cada Semana.

Nota: Se observa que, aunque la presentación de los resultados difiere respecto a los obtenidos con el paquete SAS, los valores de los estimadores, de los estadísticos y de los p del test coinciden (salvo quizá algún decimal), lo que implica que las conclusiones desprendidas de ambos análisis son idénticas.

Problema 2: Enfoques univariado y multivariado

Se desea investigar el crecimiento de la distancia entre dos puntos de la cara que interesan especialmente a los ortodoncistas en varones y mujeres. Para esto se consideraron 11 mujeres y 16 varones, y se les midió la distancia (en mm.) a los 8, 10, 12 y 14 años. Los datos encontrados fueron:

obs.	SEXO	E8	E10	E12	E14
1	1	21.0	20.0	21.5	23.0
2	1	21.0	21.5	24.0	25.5
3	1	20.5	24.0	24.5	26.0
4	1	23.5	24.5	25.0	26.5
5	1	21.5	23.0	22.5	23.5
6	1	20.0	21.0	21.0	22.5
7	1	21.5	22.5	23.0	25.0
8	1	23.0	23.0	23.5	24.0
9	1	20.0	21.0	22.0	21.5
10	1	16.5	19.0	19.0	19.5
11	1	24.5	25.0	28.0	28.0
12	2	26.0	25.0	29.0	31.0
13	2	21.5	22.5	23.0	26.5
14	2	23.0	22.5	24.0	27.5
15	2	25.5	27.5	26.5	27.0
16	2	20.0	23.5	22.5	26.0
17	2	24.5	25.5	27.0	28.5
18	2	22.0	22.0	24.5	26.5
19	2	24.0	21.5	24.5	25.5
20	2	23.0	20.5	31.0	26.0
21	2	27.5	28.0	31.0	31.5
22	2	23.0	23.0	23.5	25.0
23	2	21.5	23.5	24.0	28.0
24	2	17.0	24.5	26.0	29.5
25	2	22.5	25.5	25.5	26.0
26	2	23.0	24.5	26.0	30.0
27	2	22.0	21.5	23.5	25.5

donde Sexo 1=Femenino y Sexo 2=Masculino.

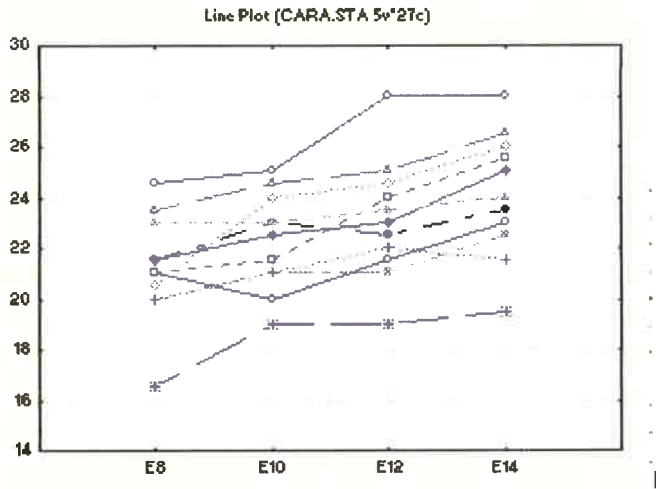
Queremos específicamente determinar si hay diferencia entre las distancias medias de varones y mujeres ("efecto Sexo").

Ingresando estos datos con el formato anterior en el paquete STATISTICA, puede examinarse el comportamiento de los datos gráficamente, según se muestra en los Gráficos 1 y 2.

Para realizar el análisis, se ingresó en primer lugar dentro de *STATISTICA Module Switcher* a *ANOVA/MANOVA*, que es el "entorno" en el que se realiza todo el trabajo.

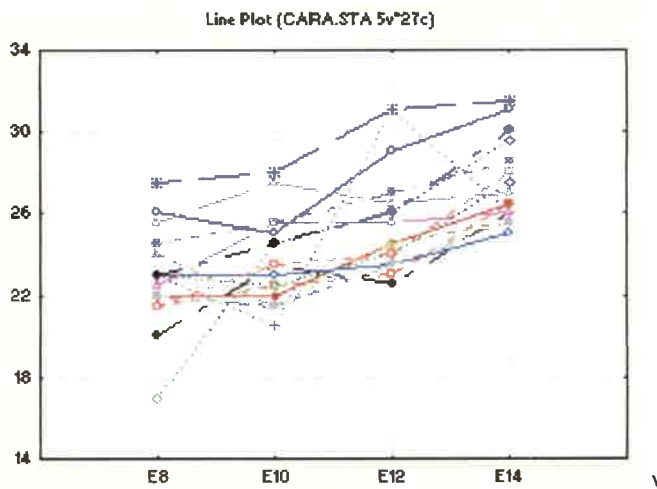
Luego, para los gráficos en sí, se utilizó dentro de *Graphs* la opción *Stats 2D Graphs*, y ahí en *Line Plots (Case Profiles)* se indicó usar las variables (*Variables*) E8, E10, E12 y E14 correspondientes a las mediciones en los cuatro tiempos, y en *Cases*, se indica para el primer gráfico *From* 1 *to* 11, y para el segundo *From* 12 *to* 27 para distinguir entre mujeres y varones. Luego, si se quieren exportar los gráficos para otras aplicaciones se puede elegir la opción *Save Bitmap*, con lo cual se graba el archivo con extensión BMP o llevarlos directamente copiando.

Gráfico 1



Mujeres

Gráfico 2



Varones

En los gráficos podemos apreciar comportamiento más homogéneo en Varones que en Mujeres, y también que los valores para Varones son mayores que para Mujeres (en conjunto).

Para indicar que se trabajará con medidas repetidas, dentro de *Analysis*, se ingresó en *Startup Panel*. Ahí se determinan las variables a usar en *Variables:* se indica como variable independiente a "Sexo" y como dependientes a "E8, E10, E12 y E14". También en *Codes for between-groups factors* con *OK* se indica tomar como niveles de la variable Sexo los indicados (1 y 2). Finalmente en *Repeated-measures (within SS) design* se indica que se tienen 4 niveles para el factor repetido al que se denomina (en este caso) Edad.

Luego aparece una ventana *ANOVA Results*, y eligiendo ahí *All-effects:*

1)

Summary of all Effects; design: (cara.sta)
1-SEXO, 2-EDAD

Effect	df Effect	MS Effect	df Error	MS Error	F	p-level
1	1	141.4120	25	15.04018	9.40228	.005146
2	3	70.3494	75	1.97071	35.69741	.000000
12	3	4.8093	75	1.97071	2.44038	.070903

Si se elige *Means/graphs* aparece el mismo cuadro anterior, y se puede elegir cualquiera de los tres renglones.

Si se elige el renglón "Sexo" y se opta por *Scrollsheet* aparece

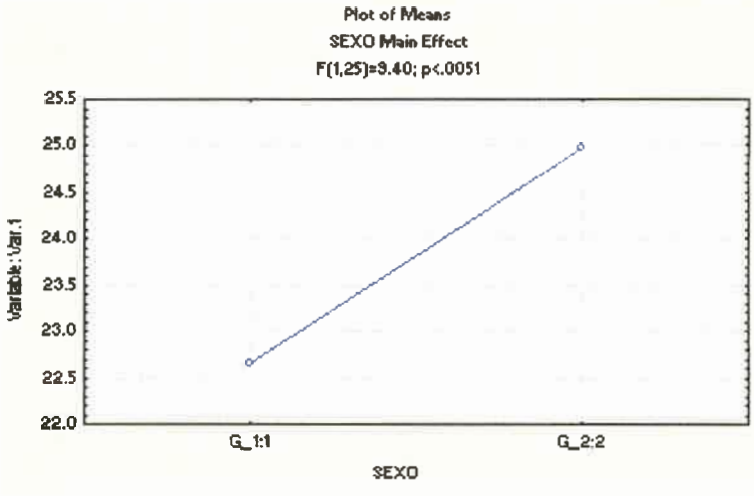
2)

Means (cara.sta)
F(1,25)=9.40; p<.0051

	Depend. Var.1
1	22.64773
2	24.97656

y eligiendo *Graphs* aparece

3)



donde se confirma la impresión dada por los gráficos iniciales (que las distancias son mayores en varones, pues aquí se vé que en promedio sí lo son).

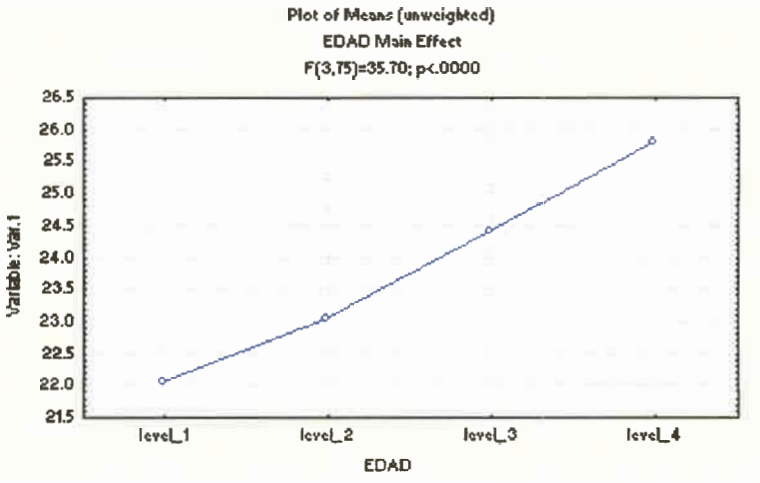
Si se elige el renglón "Edad" y se opta por *Scrollsheet* aparece

4)

```
Means (unweighted) (cara.sta)
F(3,75)=35.70; p<.0000
    Depend.
    Var.1
....        1  22.02841
....        2  23.01989
....        3  24.40483
....        4  25.79545
```

y eligiendo *Graphs* se ve

5)



donde sólo se vé que las distancias medias crecen con la edad.

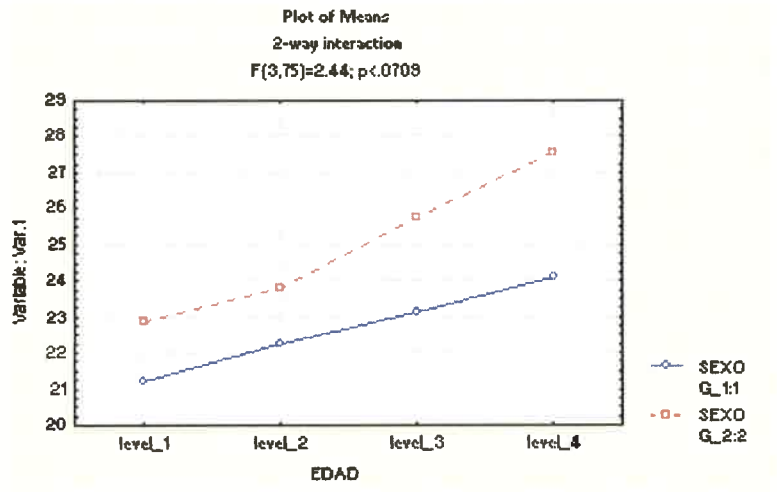
Finalmente si se elige el renglón "Sexo x Edad" y se opta por *Scrollsheet* aparece

6)

```
Means (cara.sta)
F(3,75)=2.44; p<.0709
    Depend.
    Var.1
1         1            21.18182
1         2            22.22727
1         3            23.09091
1         4            24.09091
2         1            22.87500
2         2            23.81250
2         3            25.71875
2         4            27.50000
```

y eligiendo *Graphs*

7)



donde se observa un comportamiento muy similar entre los sexos, lo que lleva a pensar la no interacción entre las variables, que también puede verse en las salidas 1) y 6) con el test de la interacción.

Otra manera de ver las tablas ANOVA para los efectos específicos, es con *Specific effect/Means/Graphs*, y al elegir el efecto Sexo aparece:

8)

```
MAIN EFFECT: SEXO (cara.sta)
1-SEXO, 2-EDAD
          Sum of                  Mean
          Squares       df        Square          F          p-level
Effect    141.4120       1      141.4120     9.402282       .005146
Error     376.0046      25       15.0402
```

más la salida de medias por sexo 2)

Al elegir el efecto Edad, como es el factor repetido, y en *Output Options* se indicó realizar los ajustes Greenhouse Geisser, el test de esfericidad y la matriz de la suma de cuadrados, aparecen las siguientes salidas:

9)

```
Mauchly's Sphericity Test (cara.sta)
MAIN EFFECT: EDAD
                        W          Chi-Sqr.          df          p
Sphericity Test      .742480       7.063498           5       .215993
```

10)

Greenhouse/Geisser & Huynh/Feldt Epsilon (cara.sta)
MAIN EFFECT: EDAD

	Epsilon
Greenhouse-Geisser Epsilon	.870279
Huynh-Feldt Epsilon	1.000000
Lower-bound Epsilon	.333333

11)

SSCP: Hypothesis (cara.sta)

	1	2	3
1	185.0038	50.5867	-47.5322
2	50.5867	13.8322	-12.9970
3	-47.5322	-12.9970	12.2122

12)

SSCP: Error (cara.sta)

	1	2	3
1	61.5795	-11.7438	-2.5179
2	-11.7438	35.3561	-14.9872
3	-2.5179	-14.9872	50.8680

13)

MAIN EFFECT: EDAD (cara.sta)
1-SEXO, 2-EDAD

	Sum of Squares	df	Mean Square	F	p-level
Effect	211.0482	3	70.34941	35.69741	.000000
Error	147.8036	75	1.97071		

14)

Univariate Test with Adjusted Degrees of Freedom (cara.sta)
F = 35.69741
MAIN EFFECT: EDAD

	Unadjstd	Greenhs. Geisser	Huynh Feldt	Lower Bound
Epsilon		.87028	1.00000	.33333
df 1	3.00000	2.61084	3.00000	1.00000
df 2	75.00000	65.27097	75.00000	25.00000
p-level	.00000	.00000	.00000	.00000

15)

MAIN EFFECT: EDAD (cara.sta)
1-SEXO, 2-EDAD

	Value	p-level
Wilks' Lambda	.19202	
Rao R Form 2 (3, 23)	32.25987	.000000
Pillai-Bartlett Trace	.80798	
V (3,23)	32.25987	.000000

más la salida 4) de medias por edad.

Al elegir los efectos Sexo y Edad conjuntamente (indicando la interacción) como en *Output Options* se indicó realizar los ajustes Greenhouse Geisser, el test de esfericidad y la matriz de sumas de cuadrados, aparecen las siguientes salidas:

16)

Mauchly's Sphericity Test (cara.sta)
INTERACTION: 1 x 2

	W	Chi-Sqr.	df	p
Sphericity Test	.742480	7.063498	5	.215993

17)

Greenhouse/Geisser & Huynh/Feldt Epsilon (cara.sta)
INTERACTION: 1 x 2

	Epsilon
Greenhouse-Geisser Epsilon	.870279
Huynh-Feldt Epsilon	1.000000
Lower-bound Epsilon	.333333

18)

SSCP: Hypothesis (cara.sta)

	1	2	3
1	9.59638	6.23762	-2.73072
2	6.23762	4.05443	-1.77496
3	-2.73072	-1.77496	.77705

19)

SSCP: Error (cara.sta)

	1	2	3
1	61.5795	-11.7438	-2.5179
2	-11.7438	35.3561	-14.9872
3	-2.5179	-14.9872	50.8680

20)

INTERACTION: 1 x 2 (cara.sta)
1-SEXO, 2-EDAD

	Sum of Squares	df	Mean Square	F	p-level
Effect	14.4279	3	4.809287	2.440377	.070903
Error	147.8036	75	1.970715		

21)

Univariate Test with Adjusted Degrees of Freedom (cara.sta)
F = 2.440377
INTERACTION: 1 x 2

	Unadjstd	Greenhs. Geisser	Huynh Feldt	Lower Bound
Epsilon		.87028	1.00000	.33333
df 1	3.00000	2.61084	3.00000	1.00000
df 2	75.00000	65.27097	75.00000	25.00000
p-level	.07090	.07222	.07090	.13082

22)

INTERACTION: 1 x 2 (cara.sta)
1-SEXO, 2-EDAD

	Value	p-level
Wilks' Lambda	.732311	
Rao R Form 2 (3, 23)	2.802480	.062517
Pillai-Bartlett Trace	.267689	
V (3,23)	2.802480	.062517

y finalmente 6) con las medias por Edad y Sexo.

Para analizar los resultados comenzamos por la Interacción, de la cual podemos ver un análisis univariado no-ajustado abreviado en 1), la ampliación en 20) y los ajustes por no esfericidad en 21). Estos últimos en realidad no serían necesarios pues con 16) y 17) (que coinciden con 9) y 10)), hemos probado que puede asumirse el cumplimiento del supuesto de esfericidad. De todas maneras el test no ajustado y los ajustados nos llevan a la misma conclusión: como p>0.05, puede decirse que *no hay evidencia de que exista*

interacción entre Sexo y Edad. A esta misma conclusión arribamos con los tests multivariados *Wilks' Lambda* y *Pillai-Bartlett Trace* ($p = 0.062517$ para ambos), que aparecen en la salida 22).

Se pueden observar entonces los factores separadamente. En el primer renglón de 1) aparece el test univariado (que es el único que se realiza), ampliado en 8) para el efecto Sexo. Como $p = 0.005146 < 0.05$ concluimos que *hay diferencia significativa entre la distancia media de Varones y de Mujeres (hay efecto Sexo).*

En cuanto a la Edad, podemos recurrir al segundo renglón de 1), que presenta el test univariado no ajustado, el cual es ampliado en 13). También puede verse éste en 14) junto con los ajustes por no esfericidad los cuales, como ya se dijo para la interacción, no serían necesarios por el cumplimiento del supuesto mostrado en 9) y 10): respecto a este supuesto, en los tests de Mauchly y de Chi-cuadrado de 9) la hipótesis nula plantea el cumplimiento del supuesto de esfericidad y como resulta $p > 0.05$ concluimos que *no hay evidencia para contradecir el cumplimiento del supuesto de esfericidad.* Esto se confirma con los estimadores de Epsilon cercanos (y para Huynh-Feldt igual) a uno que aparecen en 10). Una nueva confirmación es la similitud de los valores no ajustado y ajustados del estadístico (que aparecen en 14). Por lo tanto podemos decir que como $p = 0 < 0.05$ *hay diferencia altamente significativa entre las distancias medias por Edad (hay efecto Tiempo).* En realidad a partir de los gráficos podíamos ver cláramente ésto.

CAPITULO 4
Métodos de regresión

Ya hemos dicho que en muchas disciplinas científicas se utiliza ampliamente el análisis de la varianza, por lo que para muchos investigadores es natural usar el análisis de la varianza univariado para analizar datos provenientes de mediciones repetidas. Sin embargo, para que dicho análisis sea válido, la matriz de covarianza de las medidas repetidas debe satisfacer ciertos requerimientos (las $p-1$ variables derivadas de las p variables medidas deben satisfacer la esfericidad). Admitiendo la no esfericidad, pueden usarse pruebas F aproximadamente válidas ajustando los grados de libertad, pero se debería preferir un ajuste más natural que ese realizado luego de observar los resultados del análisis.

El análisis multivariado es una buena alternativa (como ya se dijo) cuando se sospecha (o se tiene la prueba) de la falta de esfericidad, puesto que en él no se imponen restricciones a la matriz de covarianza. Sin embargo este análisis requiere que los datos sean balanceados, en el sentido que cada individuo sea medido en las mismas p ocasiones.

Con los métodos de regresión, en principio se dejará de lado el cumplimiento del supuesto de esfericidad, permitiendo además que la matriz de covarianza sea arbitraria. Más aún, se admitirá que cada individuo sea medido en diferente número de ocasiones (en particular se podrán incluir los datos correspondientes a individuos con un conjunto incompleto de mediciones). Según Crowder y Hand (1990) el problema de observaciones perdidas o censuradas es un tópico ampliamente tratado en la literatura. Para analizar estos datos correctamente se deben tener mecanismos que las considere y permita tratarlas. Son muy comunes este tipo de observaciones en análisis de supervivencia. Algunas aplicaciones pueden verse en Gorbein at. al (1992) y en Heyting et. al (1992).

Caso especial

Una primera aproximación al tema puede realizarse de la siguiente manera: sea el vector $\boldsymbol{y}$ normal p-variado con media $\boldsymbol{\mu} = \boldsymbol{X\beta}$ y matriz de covarianza $\boldsymbol{\Sigma}$; en este caso $\boldsymbol{X}$ es una matriz pxq de variables explicatorias conocidas (indicadoras del diseño), $\boldsymbol{\beta}$ es un vector qx1 de coeficientes de regresión desconocidos y $\boldsymbol{\Sigma}$ es una matriz pxp desconocida y "sin" estructura.

Supongamos tener una muestra de n vectores de observaciones $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ de una distribución $N_p(\boldsymbol{X\beta},\boldsymbol{\Sigma})$. *Aquí con "caso especial" nos referimos a que las $\boldsymbol{y}_i$ tienen todas la misma matriz $\boldsymbol{X}$*. Cuando $\boldsymbol{\Sigma}$ es conocida, el estimador por mínimos cuadrados generalizados de $\boldsymbol{\beta}$ se obtiene minimizando la forma cuadrática $\sum_{i=1}^{n} Q_i(\boldsymbol{\beta},\boldsymbol{\Sigma})$ donde

$$Q_i(\boldsymbol{\beta},\boldsymbol{\Sigma}) = (\boldsymbol{y}_i - \boldsymbol{X\beta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{X\beta})$$

La solución es

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \overline{\boldsymbol{y}}$$

con matriz de covarianza $\boldsymbol{V}_{\boldsymbol{\beta}} = n^{-1} (\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1}$, de donde se pueden calcular los errores estándar para coeficientes individuales y pueden realizarse puebas de hipótesis para contrastes, entre otras cosas. En el caso de que $\boldsymbol{\Sigma}$ sea desconocido, se lo reemplaza por su estimador $\boldsymbol{S}$ y así resulta

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{S}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{S}^{-1} \overline{\boldsymbol{y}}$$

con matriz de covarianza estimada $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = n^{-1} (\boldsymbol{X}^T \boldsymbol{S}^{-1} \boldsymbol{X})^{-1}$. Bajo el supuesto de normalidad, éste es también el estimador de máxima verosimilitud.

Para probar la bondad del ajuste del modelo $\boldsymbol{\mu} = \boldsymbol{X\beta}$ usamos lo siguiente: la forma residual con $\boldsymbol{\mu}$ no restringido (y estimado por $\overline{\boldsymbol{y}}$) es

$$\sum_{i=1}^{n} (\boldsymbol{y}_i - \overline{\boldsymbol{y}})^T \boldsymbol{S}^{-1} (\boldsymbol{y}_i - \overline{\boldsymbol{y}}).$$

Si, en cambio, $\boldsymbol{\mu}$ está restringido a ser de la forma $\boldsymbol{X\beta}$, su estimador es $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ y el residuo es

$$\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T \boldsymbol{S}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

La diferencia entre ambos representa cuánto mejor $\overline{\boldsymbol{y}}$ ajusta los datos que $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$, y puede manipularse algebraicamente de la forma $n\ \overline{\boldsymbol{y}}^T \boldsymbol{S}^{-1}(\boldsymbol{I}_p - \boldsymbol{P})\,\overline{\boldsymbol{y}}$, donde $\boldsymbol{I}_p$ es la matriz identidad y $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{S}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{S}^{-1}$. Esta diferencia puede converstirse en un estadístico de razón de varianza, resultando un test de bondad de ajuste basado en

$$F_{p-q,n-p+q} = (n-p+q)(p-q)^{-1}\overline{\boldsymbol{y}}^T \boldsymbol{S}^{-1}(\boldsymbol{I}_p - \boldsymbol{P})\,\overline{\boldsymbol{y}}$$

Un test alternativo es uno de los dos dados por Rao(1959), basado en la minimización de la forma cuadrática

$$Q(\boldsymbol{\beta}) = (\overline{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})$$

para $\boldsymbol{\beta}$, lo cual se corresponde con elegir $\boldsymbol{\beta}$ de tal manera que $\boldsymbol{X}\boldsymbol{\beta}$ esté cerca de $\overline{\boldsymbol{y}}$. Este mínimo se alcanza en el mismo $\widehat{\boldsymbol{\beta}}$ anterior y

$$Q(\widehat{\boldsymbol{\beta}}) = \overline{\boldsymbol{y}}^T \boldsymbol{S}^{-1}(\boldsymbol{I}_p - \boldsymbol{P})\,\overline{\boldsymbol{y}}.$$

Caso general

Aquí se considera que las matrices de diseño $\boldsymbol{X}$ pueden diferir de uno a otro individuo, lo cual incluye el caso de sujetos bajo diferentes tratamientos o con valores perdidos.

Sea $\boldsymbol{y}_i = [y_{i1}, \ldots, y_{ip_i}]^T$ el vector de respuestas del i-ésimo individuo. Podemos modelar la distribución de ese vector como:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{u}_i \qquad \textbf{(4.1)}$$

donde $\boldsymbol{\beta}$ es un vector qx1 de parámetros, común a todos los individuos, que incluye los parámetros que describen tanto las diferencias entre individuos como las diferencias entre las ocasiones en que se realizan las mediciones (es

decir diferencias "dentro" de individuos); $\boldsymbol{X}_i$ es una matriz p_ixq de diseño para el individuo i-ésimo; $\boldsymbol{u}_i$ es un vector de valores aleatorios que se asume que tiene distribución normal, $\boldsymbol{u}_i \sim \mathsf{N}(\mathbf{0},\boldsymbol{\Sigma}_{p_i})$, con $\boldsymbol{\Sigma}_{p_i}$ una submatriz adecuada de la matriz $\boldsymbol{\Sigma}_p$ de covarianzas para todas las p posibles ocasiones medidas, a la que se imponen condiciones que serán resumidas más adelante.

El modelo para el análisis univariado descripto en el capítulo anterior, tiene el mismo vector de parámetros $\boldsymbol{\beta}$, pero agrupa las $\boldsymbol{y}_i$ para formar un solo vector de observaciones npx1, los $\boldsymbol{X}_i$ para formar una sola matriz de diseño npxq y los $\boldsymbol{\Sigma}_{p_i}$ para formar una sola matriz de covarianza $\boldsymbol{\Sigma}$.

El estimador general de mínimos cuadrados se obtiene minimizando la forma cuadrática $\sum_{i=1}^{n} Q_i(\boldsymbol{\beta},\boldsymbol{\Sigma})$ donde

$$Q_i(\boldsymbol{\beta},\boldsymbol{\Sigma}) = (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_{p_i}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$$

en el caso hipotético de que $\boldsymbol{\Sigma}$ sea conocido; aquí la solución es

$$\widehat{\boldsymbol{\beta}} = (\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{\Sigma}_{p_i}^{-1}\boldsymbol{X}_i)^{-1}\,(\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{\Sigma}_{p_i}^{-1}\boldsymbol{y}_i)$$

con matriz de covarianza

$$\boldsymbol{V}_{\boldsymbol{\beta}} = (\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{\Sigma}_{p_i}^{-1}\boldsymbol{X}_i)^{-1}.$$

Si $\boldsymbol{\Sigma}$ es desconocido se reemplaza $\boldsymbol{\Sigma}_{p_i}$ por su estimador $\widehat{\boldsymbol{\Sigma}}_{p_i}$ que es la submatriz apropiada de la "matriz estimador" de máxima verosimilitud $\widehat{\boldsymbol{\Sigma}}$, con lo que se obtiene

$$\widehat{\boldsymbol{\beta}} = (\sum_{i=1}^{n}\boldsymbol{X}_i^T\widehat{\boldsymbol{\Sigma}}_{p_i}^{-1}\boldsymbol{X}_i)^{-1}\,(\sum_{i=1}^{n}\boldsymbol{X}_i^T\widehat{\boldsymbol{\Sigma}}_{p_i}^{-1}\boldsymbol{y}_i) \qquad \textbf{(4.2)}$$

Nuevamente, bajo normalidad, éste es también el estimador que se obtiene por el método de máxima verosimilitud usando la función de verosimilitud

$$L = -\tfrac{1}{2}\Big\{\sum_i log|\boldsymbol{\Sigma}_{p_i}| - \sum_i(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_{p_i}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\Big\}$$

que a su vez se basa en (4.1).

Cuando no hay valores perdidos

$$\widehat{\boldsymbol{\Sigma}} = n^{-1} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}})(\boldsymbol{y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}})^T.$$

Cuando sí los hay, las ecuaciones se vuelven muy difíciles de manejar, y por esto se suele utilizar la fórmula usual para la matriz, con $\widehat{\boldsymbol{\Sigma}}_{jk}$ el promedio de $(\boldsymbol{y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}})_j (\boldsymbol{y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}})_k$ sobre todos los valores de la muestra, pero promediando sólo sobre los componentes presentes para cada combinación (j,k).

En general se prefiere el divisor $n - q$ a n para estimar $\boldsymbol{\Sigma}$, pues así se tiene en cuenta la pérdida de grados de libertad al estimar $\boldsymbol{\beta}$ y se obtienen estimadores insesgados en casos balanceados. Usualmente no se pueden obtener soluciones explícitas para $\widehat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\Sigma}}$ separadamente y las ecuaciones deben resolverse iterativamente. Hay soluciones aproximadas para casos particulares, pero ese no es tema de la presente monografía. Además, desafortunadamente, el test F de bondad de ajuste (dado para el caso particular antes presentado) no es ya adecuado.

La matriz de covarianza estimada de $\widehat{\boldsymbol{\beta}}$ es

$$\widehat{\boldsymbol{V}}_{\beta} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i^T \widehat{\boldsymbol{\Sigma}}_{p_i}^{-1} \boldsymbol{X}_i \right)^{-1}$$

que puede usarse en tests de hipótesis e intervalos de confianza de la manera usual. Por ejemplo, para la hipótesis

$$\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$$

con $\boldsymbol{H}$ una matriz rxq y $\boldsymbol{h}$ un vector rx1 especificados. El estadístico de contraste es

$$(\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h})^T (\boldsymbol{H}\widehat{\boldsymbol{V}}_{\beta}\boldsymbol{H}^T)^{-1} (\boldsymbol{H}\widehat{\boldsymbol{\beta}} - \boldsymbol{h})$$

que tiene aproximadamente distribución χ^2_r para tamaños de muestra grandes.

Ejemplo 4.1: Supongamos tener dos grupos de individuos medidos cada uno en los tiempos 1, 2 y 3 (ver Ejemplo 2.4). Modelamos la respuesta esperada del k-ésimo grupo como

$$\begin{bmatrix} \mu_{k1} + \mu_{k2} & \mu_{k1} + \mu_{k3} & \mu_{k1} - \mu_{k2} - \mu_{k3} \end{bmatrix}^T$$

Entonces para un individuo del primer grupo que es medido en las tres ocasiones tenemos

$$\mathrm{E}\begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix}$$

y para un individuo en el segundo grupo es

$$\mathrm{E}\begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix}$$

Para un individuo del primer grupo que sólo fue medido en los tiempos 1 y 3 tendríamos

$$\mathrm{E}\begin{pmatrix} y_{i1} \\ y_{i3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix}$$

Ya se habían mencionado las restricciones a la forma de la matriz de covarianza Σ_p. Ésta es descripta frecuentemente en términos de un pequeño conjunto de parámetros, lo cual puede llevar a estimadores "improvisados" si

dicho número es mucho menor que el número de términos distintos en Σ_p (que conducen a estructuras naturales para Σ_p). Ejemplos de estructuras de este tipo son:

1 - $\Sigma_p = \sigma^2 I$ que surge si todas las observaciones pueden ser consideradas como independientes con idéntica varianza y un solo parámetro.

2 - El modelo de simetría compuesta (visto en el capítulo anterior), donde $\Sigma_p = \sigma_a^2 \mathbf{1}\mathbf{1}^T + \sigma^2 I$ con dos parámetros.

3 - El modelo de efectos aleatorios cuyo número de parámetros depende del número de efectos aleatorios.

4 - Modelos que parametrizan correlaciones entre términos de error sucesivos, más allá de cualquier correlación inducida por efectos aleatorios entre individuos.

5 - Modelos que combinan 3 y 4.

En general, se considera que

$$\Sigma_{p_i} = \Sigma_{p_i}(\phi)$$

donde ϕ denota el vector de parámetros que restringen a Σ_p y es asumido como el mismo para todos los individuos. La log-verosimilitud es así asumida como una función de β y de ϕ. El estimador de máxima verosimiltud de β es así una función de ϕ. Si lo sustituímos en la expresión general de verosimilitud, obtenemos la solución maximizando sobre todo ϕ.

Comparación de líneas de regresión

Variables indicadoras

Si bien las variables utilizadas en el análisis de regresión son generalmente de tipo cuantitativo, en muchos casos es necesario usar variables cualitativas o categorizadas como predictoras (tratamiento, raza, status del empleado, sexo, turno de trabajo, etc.). Como por lo general una variable cualitativa no tiene una escala natural de medida, se deben asignar

niveles a los "valores" de la variable para determinar el efecto que ésta podría tener sobre la respuesta. Esto se logra usando variables indicadoras.

Ejemplo 4.2:: Supongamos que se aplican 2 tratamientos, cada uno a cinco individuos y se mide cierta variable respuesta en varios tiempos (en el Problema 3 resolveremos una situación similar). Queremos determinar si los perfiles de respuesta son iguales en ambos grupos.

Observemos que en este caso queremos relacionar la variable respuesta (y) con el tiempo y el tratamiento. La segunda variable regresora es de tipo cualitativa con dos niveles (A y B). Usaremos una variable indicadora que tome los valores 0 y 1 para indicar las categorías de la variable "tratamiento".

Sea $x_2 = 0$ si se recibió el tratamiento A
1 si se recibió el tratamiento B

La elección de 0 y 1 para identificar los niveles de la variable son arbitrarios. Cualquier par de valores distintos pueden ser utilizados, aunque 0 y 1 son generalmente los mejores por simplificar notablemente los cálculos.

Asumiendo que un modelo lineal es apropiado, si no consideramos los subíndices correspondientes a tratamiento y sujeto, tenemos

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \qquad \textbf{(4.3)}$$

Para interpretar los parámetros en este modelo, consideremos primero el tratamiento A, para el cual $x_2 = 0$. El modelo para este caso queda:

$$y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \epsilon = \beta_0 + \beta_1 x_1 + \epsilon \qquad \textbf{(4.4)}$$

que es la relación entre la respuesta y y el tiempo x_1 para el tratamiento A. Ésta es una recta con pendiente β_1 y ordenada al origen β_0.

Para el tratamiento B, se tiene $x_2 = 1$ y el modelo queda:

$$y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \epsilon = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon \qquad \textbf{(4.5)}$$

que es la relación entre la respuesta y y el tiempo x_1 para el tratamiento B. Ésta es una recta con pendiente β_1 y ordenada al origen $\beta_0 + \beta_2$.

Los modelos (4.4) y (4.5) describen dos líneas de regresión paralelas, es decir, dos líneas con pendiente común β_1 y diferentes ordenadas al origen. Asumimos, como es usual, que la varianza de los errores ϵ es la misma para ambos tratamientos. El parámetro β_2 expresa la diferencia en altura entre las dos rectas de regresión, es decir es una medida de la diferencia entre la respuesta media del tratamiento A y el B.

Si se tuvieran más de dos niveles en la variable categórica, se puede generalizar esta aproximación de la siguiente manera: si se está interesado en comparar tres tratamientos A, B y C, se necesitan dos variables indicadoras para incorporar al modelo los tres niveles de tratamiento. Los niveles de las variables indicadoras son:

x_2	x_3	
0	0	si la observación corresponde al tratamiento A
1	0	si la observación corresponde al tratamiento B
0	1	si la observación corresponde al tratamiento C

y el modelo de regresión es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \qquad \textbf{(4.6)}$$

En general una variable cualitativa con a niveles es representada por $a - 1$ variables indicadoras, cada una con valores 0 y 1.

Volviendo al Ejemplo 4.2, si por alguna razón se espera que las dos líneas de regresión de y con el tiempo difieran tanto en ordenada al origen como en pendiente, se puede modelar la situación con una sola ecuación de regresión utilizando variables indicadoras. El modelo es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \qquad \textbf{(4.7)}$$

Si se compara ésta con (4.3), se observa que se ha agregado un producto cruzado entre el tiempo (x_1) y la variable indicadora (x_2) al modelo.

Nuevamente para interpretar los parámetros del modelo consideremos primero el tratamiento A, para el cual resulta

$$y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3 x_1(0) + \epsilon = \beta_0 + \beta_1 x_1 \qquad \textbf{(4.8)}$$

que es una recta con ordenada al origen β_0 y pendiente β_1 (como antes).

En cambio para el tratamiento B, tenemos

$$y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3 x_1(1) + \epsilon = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1 + \epsilon \qquad \textbf{(4.9)}$$

que es también una recta pero con ordenada al origen $\beta_0 + \beta_2$ y pendiente $\beta_1 + \beta_3$.

El parámetro β_2 refleja, como antes, el cambio en la altura asociado con el cambio del tratamiento A al B, y β_3 indica el cambio en la pendiente asociado con el cambio del tratamiento A al B.

Ajustar el modelo (4.7) es equivalente a ajustar dos modelos por separado. La ventaja del uso de variables indicadoras es que los tests de hipótesis pueden desarrollarse directamente usando el llamado *método de la suma de cuadrados extra*.

Por ejemplo, sean idénticos o no los dos modelos de regresión, podemos probar

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \beta_2 \neq 0 \quad \text{y/o} \quad \beta_3 \neq 0$$

Si H_0 no es rechazada, podría indicar que un único modelo puede explicar la relación entre y y el tiempo x_1.

Si se rechaza H_0 se concluiría que las rectas difieren en pendiente y/o en ordenada al origen, y se debería continuar. Para probar si las dos rectas de

regresión tienen una pendiente común pero posiblemente distinta ordenada al origen, planteamos

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

Usando el modelo (4.7), tanto las líneas de regresión ajustadas como los tests correspondientes pueden calcularse utilizando un programa de computadora que provea las sumas de cuadrados $SC_R(\beta_1|\beta_0)$, $SC_R(\beta_2|\beta_0,\beta_1)$ y $SC_R(\beta_3|\beta_0,\beta_1,\beta_2)$, las cuales se verán en detalle en los problemas resueltos.

Lo que sí destacamos aquí es la fórmula de cálculo de los estadísticos de los tests anteriores:

1) Para el test que postula $H_0 : \beta_2 = \beta_3 = 0$, es:

$$F = \frac{SC_R(\beta_2, \beta_3|\beta_0,\beta_1)/2}{CM_E}$$

que tiene distribución F con los grados de libertad del numerador y el denominador.

2) Para el test que prueba $H_0 : \beta_3 = 0$, es:

$$F = \frac{SC_R(\beta_3|\beta_0,\beta_1, \beta_2)/1}{CM_E}$$

también con distribución F con los grados de libertad del numerador y el denominador.

Nota: Los valores 2 y 1 que están en los denominadores de los numeradores de los estadísticos dependen del problema, pero son los adecuados para el Ejemplo 4.2, que se resolverá en el Problema 3.

Las variables indicadoras son útiles en muchas situaciones de regresión. Para ver casos de variables indicadoras con más de dos niveles en los que se necesitan dos o más variables indicadoras diferentes, y comparación de M modelos de regresión diferentes remitirse a Montgomery y Peck (1982).

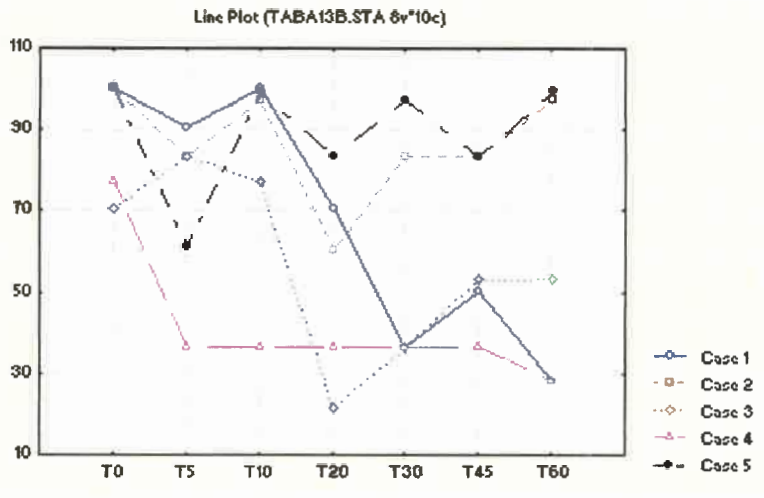
Problema 3

En un experimento médico interesa comparar 2 tratamientos para curar cierta afección. Para esto se tomaron 10 sujetos y se le aplicó el tratamiento A a 5 de ellos y el B a los restantes. A cada uno de los sujetos se les midió la variable respuesta y luego de transcurridos 0, 5, 10, 20, 30, 45 y 60 minutos desde que se aplicó el tratamiento. Se quiere determinar si los dos tratamientos provocan iguales perfiles de respuesta.

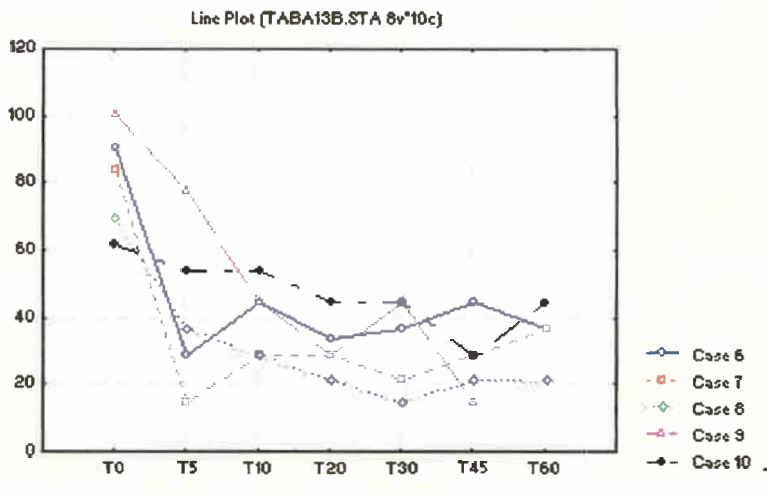
Los datos encontrados experimentalmente fueron:

Grupo	Suj	Tiempo	Y	Grupo	Suj	Tiempo	Y
1	1	0	100	2	6	0	90
1	1	5	90	2	6	5	28
1	1	10	100	2	6	10	44
1	1	20	70	2	6	20	33
1	1	30	36	2	6	30	36
1	1	45	50	2	6	45	44
1	1	60	28	2	6	60	36
1	2	0	100	2	7	0	83
1	2	5	83	2	7	5	14
1	2	10	97	2	7	10	28
1	2	20	60	2	7	20	28
1	2	30	83	2	7	30	21
1	2	45	83	2	7	45	28
1	2	60	97	2	7	60	36
1	3	0	70	2	8	0	69
1	3	5	83	2	8	5	36
1	3	10	77	2	8	10	28
1	3	20	21	2	8	20	21
1	3	30	36	2	8	30	14
1	3	45	53	2	8	45	21
1	3	60	53	2	8	60	21
1	4	0	77	2	9	0	100
1	4	5	36	2	9	5	77
1	4	10	36	2	9	10	44
1	4	20	36	2	9	20	28
1	4	30	36	2	9	30	44
1	4	45	36	2	9	45	14
1	4	60	28	2	9	60	.
1	5	0	100	2	10	0	61
1	5	5	61	2	10	5	53
1	5	10	99	2	10	10	53
1	5	20	83	2	10	20	44
1	5	30	97	2	10	30	44
1	5	45	83	2	10	45	28
1	5	60	100	2	10	60	44

Como en los problemas anteriores en primer lugar realizamos gráficos de los datos para cada grupo



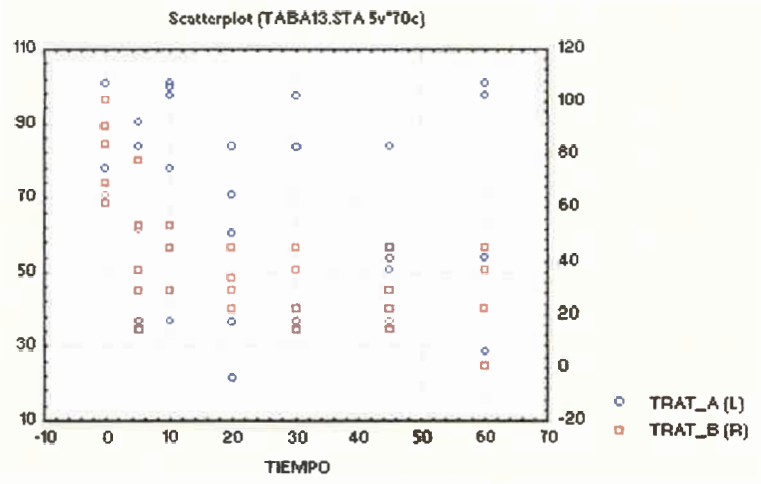
Trat. A



Trat. B

Estos gráficos sugieren que podría pensarse en una regresión lineal para realizar el ajuste.

Otro gráfico que suele realizarse es el que figura a continuación, que muestra que los datos no se presentan como un único "conglomerado", sino que se deberían rectas de regresión con distinta pendiente y distinta ordenada al origen.



Así, ajustaremos el modelo dado en (4.7)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \qquad (1)$$

que, como se dijo, al contener el producto cruzado entre el tiempo (x_1) y la variable indicadora (x_2), lleva a los modelos

$$y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3 x_1(0) + \epsilon = \beta_0 + \beta_1 x_1 \qquad (2)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3 x_1(1) + \epsilon = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon \qquad (3)$$

para el tratamiento A y B respectivamente, los cuales son rectas con distinta pendiente y distinta ordenada al origen.

El programa en SAS para ajustar el modelo completo es:

```
data gaby;
infile 'taba13.prn';
input grupo$ y tiempo ind;
proc glm;
model y=tiempo ind tiempo*ind ;
run;
```

Que da por resultado lo siguiente:

1)

```
                Number of observations in data set = 70
NOTE: Due to missing values, only 69 observations can be used in this analysis.
Dependent Variable: Y
                                  Sum of                Mean
Source                    DF     Squares              Square      F Value      Pr > F
Model                      3   18739.368            6246.456        12.12      0.0001
Error                     65   33490.284             515.235
Corrected Total           68   52229.652

                  R-Square             C.V.      Root MSE                Y Mean
                  0.358788        41.53319        22.699                54.652

Source                    DF       Type I SS     Mean Square      F Value      Pr > F
TIEMPO                     1        5591.423        5591.423        10.85      0.0016
IND                        1       12996.795       12996.795        25.22      0.0001
TIEMPO*IND                 1         151.149         151.149         0.29      0.5899
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TIEMPO	1	2280.3493	2280.3493	4.43	0.0393
IND	1	4132.0433	4132.0433	8.02	0.0062
TIEMPO*IND	1	151.1494	151.1494	0.29	0.5899

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	77.53838631	13.01	0.0001	5.96026595
TIEMPO	-0.39511002	-2.10	0.0393	0.18781060
IND	-23.96316507	-2.83	0.0062	8.46183367
TIEMPO*IND	-0.14736785	-0.54	0.5899	0.27208353

Para analizar los residuos de este modelo, debe añadirse al programa luego del comando MODEL, lo siguiente:

```
output r=resid p=predic;
proc univariate plot normal;
var resid;
proc plot;
plot resid*predic=grupo vpos=20 hpos=50;
plot resid*teimpo=grupo vpos=20 hpos=50;
```

que da por resultado, entre otras cosas:

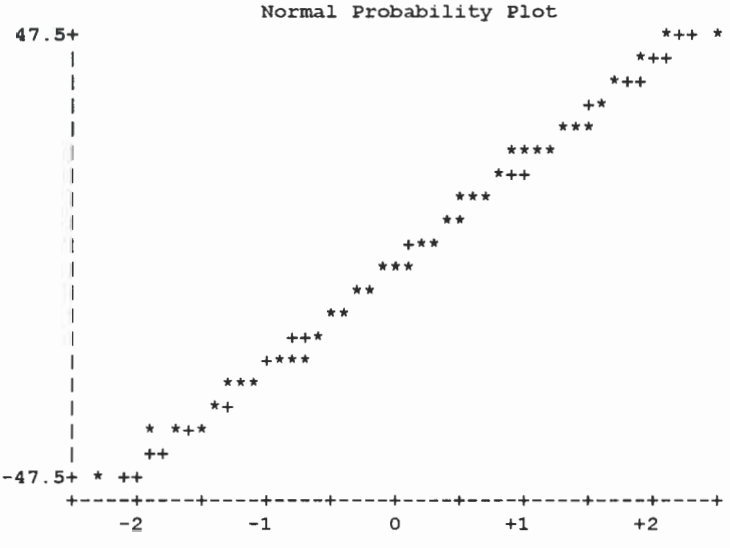
2)

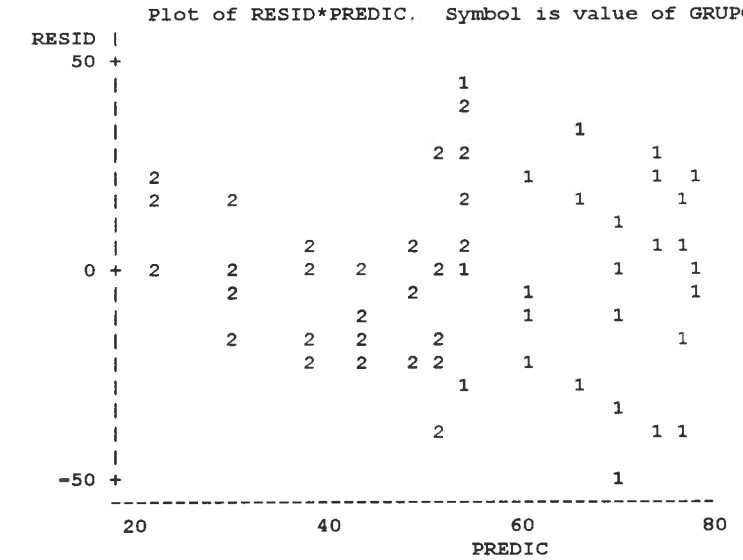
Variable=RESID

Moments

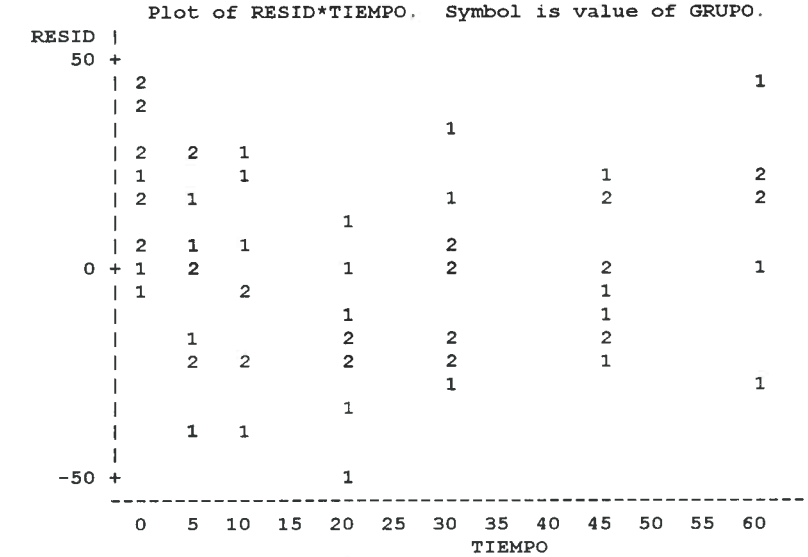
N	69	Sum Wgts	69
Mean	0	Sum	0
Std Dev	22.19244	Variance	492.5042
Skewness	0.035834	Kurtosis	-0.62131
USS	33490.28	CSS	33490.28
CV	.	Std Mean	2.671655
T:Mean=0	0	Pr>|T|	1.0000
Num ^= 0	69	Num > 0	33
M(Sign)	-1.5	Pr>=|M|	0.8099
Sgn Rank	-13.5	Pr>=|S|	0.9364
W:Normal	0.974499	Pr<W	0.4026

Normal Probability Plot





NOTE: 1 obs had missing values. 16 obs hidden.



NOTE: 1 obs had missing values. 22 obs hidden.

En líneas generales, tanto los gráficos como el test de normalidad no muestran apartamiento de los supuestos.

En cuanto a la salida 1), la Suma de Cuadrados del Modelo es la que se usa en los tests bajo la denominación de $SC_R(\beta_1,\beta_2,\ \beta_3|\beta_0)$, y el Cuadrado Medio del Error es el que allá se usa como CM_E. También ahí pueden leerse los estimadores de los parámetros del modelo.

A partir de ahí el modelo de regresión (1) estimado es:

$$\widehat{y} = 77,53838631 - 0,39511002x_1 - 23,96316507x_2 - 0,14736785x_1x_2$$

y de ahí los modelos (2) y (3) estimados son:

para el Tratamiento A:

$$\widehat{y} = 77,53838631 - 0,39511002x_1$$

para el Tratamiento B:

$$\widehat{y} = (77,53838631 - 23,96316507) + (-0,39511002 - 0,14736785)x_1$$
$$= 53,57522124 - 0,53947787x_1$$

En cuanto a los tests para β_2 y β_3 (que son los que permiten determinar si las rectas pueden asumirse como iguales en ambos tratamientos) se ejecuta dos veces el primer programa de SAS mostrado cambiando el comando MODEL por:

```
model y=tiempo ind ;
```

que da por resultado:

3)

```
Dependent Variable: Y
                                 Sum of          Mean
Source                    DF     Squares        Square       F Value     Pr > F
Model                      2   18588.219      9294.109         18.23     0.0001
Error                     66   33641.434       509.719
Corrected Total           68   52229.652

                  R-Square         C.V.       Root MSE            Y Mean
                  0.355894     41.31025         22.577            54.652

Source                    DF     Type I SS     Mean Square       F Value     Pr > F
TIEMPO                     1    5591.423      5591.423          10.97     0.0015
IND                        1   12996.795     12996.795          25.50     0.0001

Source                    DF     Type I SS     Mean Square       F Value     Pr > F
TIEMPO                     1    6041.423      6041.423          11.85     0.0010
IND                        1   12996.795     12996.795          25.50     0.0001

                                          T for H0:     Pr > |T|     Std Error of
Parameter            Estimate             Parameter=0                   Estimate
INTERCEPT         79.24364129                   15.74       0.0001      5.03370374
TIEMPO            -0.46532641                   -3.44       0.0010      0.13516161
IND               -27.46105715                  -5.05       0.0001      5.43831543
```

que es el ajuste del modelo sin interacción, y donde puede leerse en la Suma de Cuadrados del Modelo la denominada $SC_R(\beta_1,\beta_2|\beta_0)$.

Si se usa el comando:

```
model y=tiempo ;
```

se obtiene:

4)

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5591.4233	5591.4233	8.03	0.0061
Error	67	46638.2288	696.0930		
Corrected Total	68	52229.6522			

R-Square	C.V.	Root MSE	Y Mean
0.107055	48.27543	26.384	54.652

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TIEMPO	1	5591.4233	5591.4233	8.03	0.0061

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TIEMPO	1	5591.4233	5591.4233	8.03	0.0061

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	65.28860628	13.28	0.0001	4.91656942
TIEMPO	-0.44750843	-2.83	0.0061	0.15789692

que es el ajuste del modelo sin la variable indicadora donde puede leerse en la Suma de Cuadrados del Modelo la denominada $SC_R(\beta_1|\beta_0)$.

Así, en base a las salidas anteriores, podemos llevar a cabo en primer lugar el test cuyas hipótesis son:

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \beta_2 \neq 0 \quad \text{y/o} \quad \beta_3 \neq 0$$

que postula en la hipótesis alternativa que las rectas para ambos tratamientos difieren en ordenada al origen y/o en pendiente. El estadístico para este test es

$$F = \frac{SC_R(\beta_2,\, \beta_3|\beta_0,\beta_1)\,/2}{CM_E}$$

con distribución $F_{2,65}$(2 se obtiene de los grados de libertad de $SC_R(\beta_2,\, \beta_3|\beta_0,\beta_1)$ y 65 de los del Cuadrado Medio del Error que es el denominador del estadístico).

Para obtener el valor numérico del estadístico, debe calcularse $SC_R(\beta_2,\, \beta_3|\beta_0,\beta_1)$ en base a las sumas de cuadrados obtenidas de la salida, de la siguiente manera:

$$\begin{aligned} SC_R(\beta_2,\, \beta_3|\beta_0,\beta_1) &= SC_R(\beta_1,\beta_2,\, \beta_3|\beta_0) - SC_R(\beta_1|\beta_0) \\ &= 18739,368 - 5591,4233 \\ &= 13147,9447 \end{aligned}$$

Por otra parte el valor 2 del numerador de F, no es constante sino que se obtiene restando los grados de libertad de las sumas de cuadrados que intervienen en el cálculo anterior (en este caso $3-1$). Cabe señalar que $SC_R(\beta_1|\beta_0)$ como corresponde a la "debida al tiempo", puede no sólo leerse de la salida 4), sino también de la 1) en la Suma de Cuadrados tipo I de Tiempo. Así

$$F = \frac{13147,9447/2}{515,235} = 12,7591$$

que, al compararse con $F_{0.95;2,65} = 3,15$, lleva a rechazar la hipótesis nula, por lo cual se concluye que las rectas de regresión no son idénticas.

Al seguir adelante se plantea

$$\begin{aligned} \mathsf{H}_0 &: \beta_3 = 0 \\ \mathsf{H}_1 &: \beta_3 \neq 0 \end{aligned}$$

cuya hipótesis alternativa postula que las rectas para ambos tratamientos tienen distinta pendiente. El estadístico para este test es

$$F = \frac{SC_R(\beta_3|\beta_0,\,\beta_1,\,\beta_2)/1}{CM_E}$$

con distribución $F_{1,65}$ (con 1 obtenido de los grados de libertad de $SC_R(\beta_3|\beta_0,\beta_1,\beta_2)$). Este cálculo es:

$$\begin{aligned} SC_R(\beta_3|\beta_0,\beta_1,\beta_2) &= SC_R(\beta_1,\beta_2,\beta_3|\beta_0) - SC_R(\beta_1,\beta_2|\beta_0) \\ &= 18739,368 - 18588,219 \\ &= 151.149 \end{aligned}$$

de donde

$$F = \frac{151.149}{515,235} = 0,29$$

que, al compararse con $F_{0.95;1,65} = 4$, lleva a no rechazar la hipótesis nula, por lo cual no puede decirse que las rectas tienen distinta pendiente, entonces en conjunción con el test anterior, podemos afirmar:

Las rectas correspondientes a los dos tratamientos difieren en ordenada al origen y no en pendiente. (Esto podía verse directamente con un test t para el parámetro β_3, p = 0,5899 y β_2, p = 0,0062 de la salida 1)).

Por ello debería ajustarse es un modelo sin interacción, es decir

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{4}$$

que lleva a los modelos

$$y = \beta_0 + \beta_1 x_1 + \beta_2(0) + \epsilon = \beta_0 + \beta_1 x_1 \tag{5}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \epsilon = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon \tag{6}$$

para el tratamiento A y B respectivamente, los cuales son rectas con igual pendiente y distinta ordenada al origen.

Cabe señalar que (5) coincide con (2), es decir que la recta para el tratamiento A es igual con ambos modelos.

Ajustando este modelo resulta la salida 3) anterior, de donde la recta estimada es:

$$\widehat{y} = 79,24364129 - 0,46532641 x_1 - 27,46105715 x_2$$

y de ahí los modelos (5) y (6) estimados son:

para el Tratamiento A:

$$\widehat{y} = 79,24364129 - 0,46532641 x_1$$

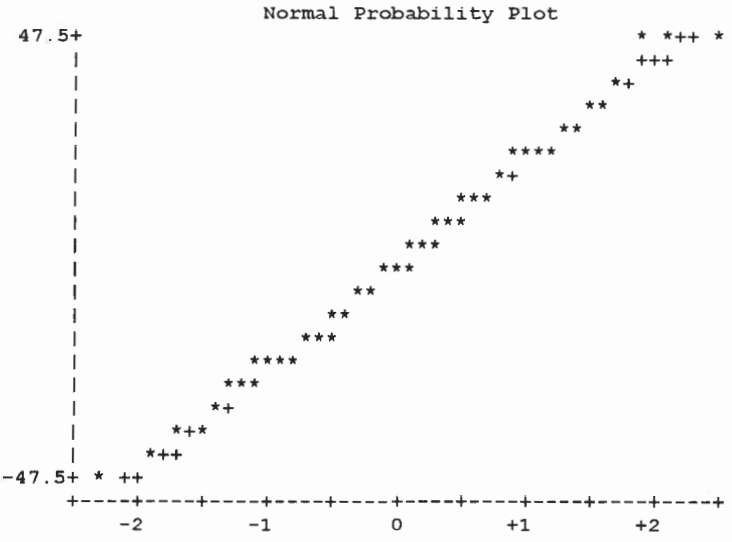
para el Tratamiento B:

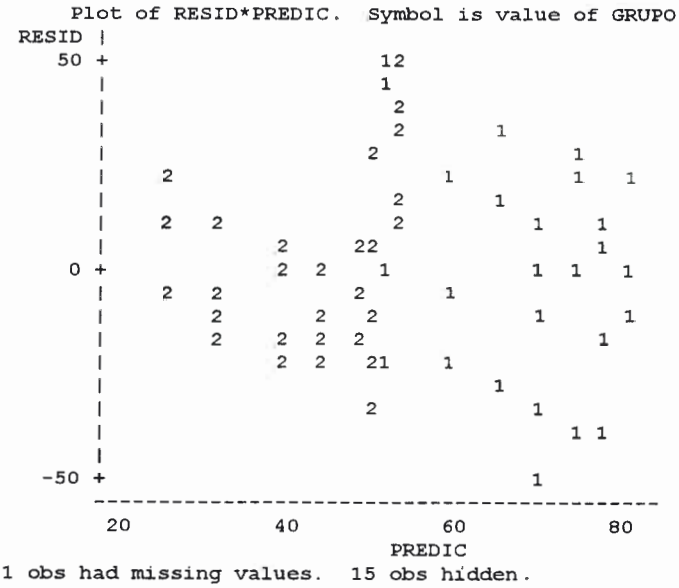
$$\begin{aligned}\widehat{y} &= (79,24364129 - 27,46105715) - 0,46532641 x_1 \\ &= 51,78258414 - 0,46532641 x_1\end{aligned}$$

que son rectas paralelas.

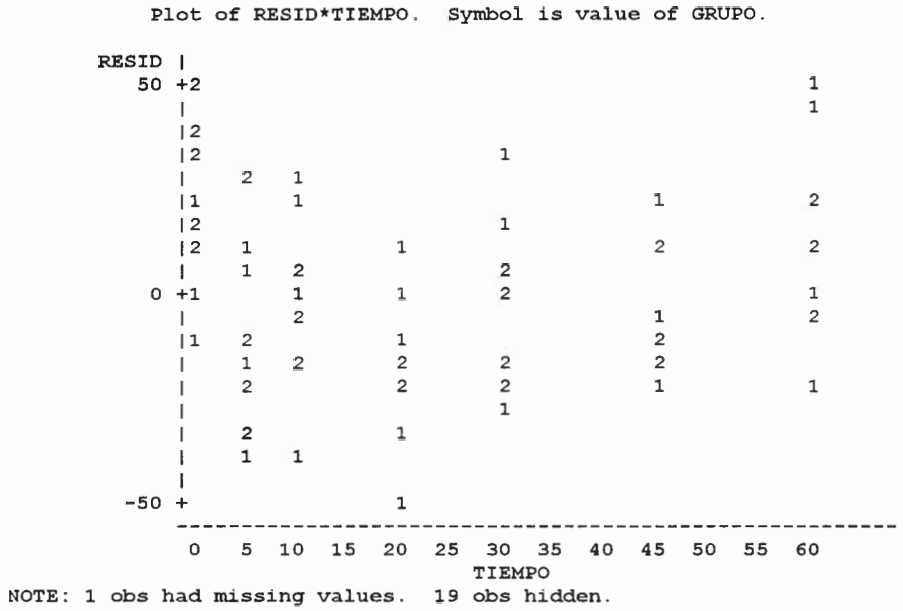
Para finalizar se muestran los gráficos y el test de normalidad de los supuestos del modelo, los cuales parecen cumplirse aproximadamente

```
                General Linear Models Procedure
              Number of observations in data set = 70

Univariate Procedure
Variable=RESID                  Moments

                N                  69     Sum Wgts            69
                Mean                0     Sum                  0
                Std Dev      22.24246     Variance        494.727
                Skewness     0.109536     Kurtosis       -0.4481
                USS          33641.43     CSS            33641.43
                CV                .      Std Mean       2.677677
                T:Mean=0            0     Pr>|T|           1.0000
                Num ^= 0           69     Num > 0            34
                M(Sign)          -0.5     Pr>=|M|          1.0000
                Sgn Rank        -14.5     Pr>=|S|          0.9317
                W:Normal      0.97891     Pr<W             0.5874
```





NOTE: 1 obs had missing values. 15 obs hidden.

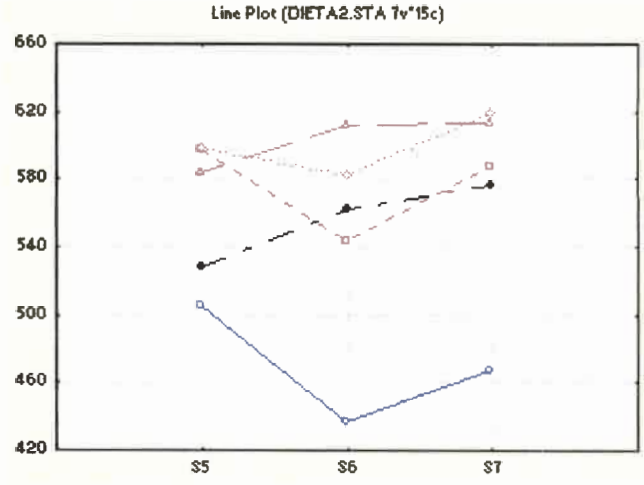


NOTE: 1 obs had missing values. 19 obs hidden.

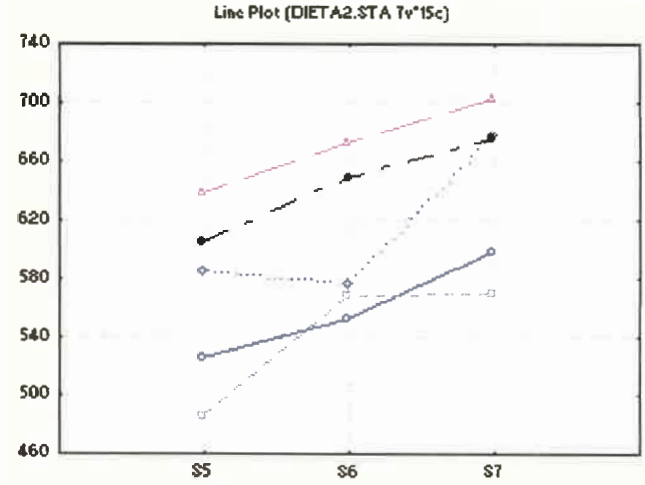
Problema 4

Usando parte de los datos del Problema 1, específicamente los correspondientes a las Semanas denotadas 5, 6 y 7 (que son en realidad las semanas 1, 2 y 3 después de haber administrado el tratamiento), nuestro interés radica principalmente en determinar si los tres grupos tienen distintas tazas de crecimiento.

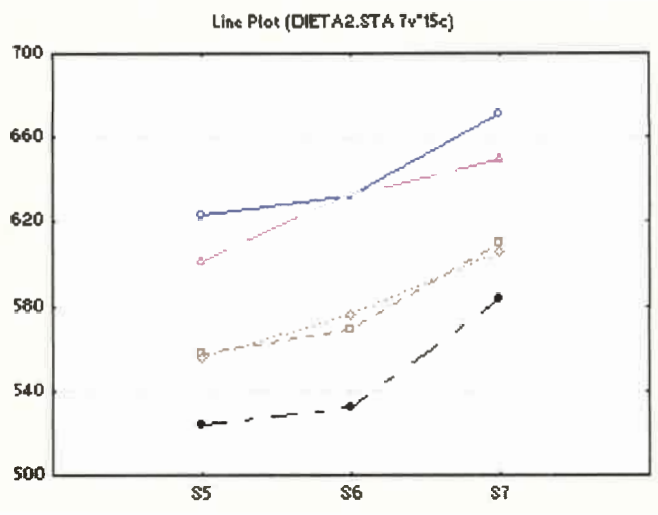
Como en los problemas anteriores en primer lugar realizamos gráficos de los datos para cada grupo



Grupo 1



Grupo 2



Grupo 3

Estos gráficos sugieren que podría pensarse en una regresión lineal para realizar el ajuste.

Otro gráfico que suele realizarse es el siguiente.

```
          Plot of PESO*SEMANA.  Symbol is value of GRUPO.

PESO |
 700 +                                                    2
     |                                                    2
     |                                 2                  3
     |                                 2                  3
     |      2                          3
     |      3                          1                  1
 600 +      1                                             2
     |      1                          1                  1
     |                                 1                  2
     |      3                          1
     |      1                          3
     |      2
 500 +      1
     |      2
     |                                                    1
     |
     |                                 1
     |
 400 +
     ------------------------------------------------------------
           1                         2                         3
                                  SEMANA
NOTE: 18 obs hidden.
```

Esto reafirma la idea de que una regresión lineal ajustaría adecuadamente estos datos, aunque los grupos aparecen bastante "mezclados" por lo que tal vez una sola línea sirva para predecir (los perfiles de crecimiento serían en ese caso los mismos para los tres grupos).

Para realizar el análisis estadístico usando variables indicadoras, como se tienen 3 niveles en la variable Grupo, se necesitan 2 variables indicadoras:

x_2	x_3	
0	0	si la observación corresponde al grupo 1
1	0	si la observación corresponde al grupo 2
0	1	si la observación corresponde al grupo 3

Si consideramos inicialmente que tanto las ordenadas al origen como las pendientes difieren al cambiar de grupo, el modelo de regresión es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon \qquad (1)$$

de donde resultan los siguientes modelos para cada Grupo:

Grupo 1: $$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) + \beta_4(x_1.0) + \beta_5(x_1.0) + \epsilon \\ &= \beta_0 + \beta_1 x_1 + \epsilon \end{aligned}$$

Grupo 2: $$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0) + \beta_4(x_1.1) + \beta_5(x_1.0) + \epsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1 + \epsilon \end{aligned}$$

Grupo 3: $$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1) + \beta_4(x_1.0) + \beta_5(x_1.1) + \epsilon \\ &= (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1 + \epsilon \end{aligned}$$

observando los modelos, vemos que los parámetros β_2 y β_3 representan cambios en la ordenada al origen, mientras que β_4 y β_5 representan cambios en la pendiente.

Para ajustar el modelo (1) usando el paquete SAS se realizó el siguiente programa, que además asigna los valores 0 y 1 a las variables indicadoras (ahí denominadas "ind1" e "ind2").

```
data gaby;
infile 'a:ta12reg.prn';
input grupo sujeto semana peso;
ind1=0;
ind2=0;
if grupo=2 then ind1=1;
if grupo=3 then ind2=1;
proc glm;
  model peso=semana ind1 ind2 semana*ind1 semana*ind2 ;
run;
```

La salida de éste es:

1)

```
                Number of observations in data set = 45

Dependent Variable: PESO
                                 Sum of          Mean
Source                DF        Squares        Square      F Value      Pr > F
Model                  5      38064.511      7612.902         2.94      0.0240
Error                 39     101012.067      2590.053
Corrected Total       44     139076.578
```

R-Square	C.V.	Root MSE	Y Mean
0.273695	8.679143	50.893	586.38

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SEMANA	1	16054.533	16054.533	6.20	0.0172
IND1	1	7525.878	7525.878	2.91	0.0962
IND2	1	8840.833	8840.833	3.41	0.0723
SEMANA*IND1	1	3542.017	3542.017	1.37	0.2493
SEMANA*IND2	1	2101.250	2101.250	0.81	0.3733

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SEMANA	1	270.4000	270.4000	0.10	0.7483
IND1	1	518.5714	518.5714	0.20	0.6570
IND2	1	47.6190	47.6190	0.02	0.8928
SEMANA*IND1	1	5544.4500	5544.4500	2.14	0.1515
SEMANA*IND2	1	2101.2500	2101.2500	0.81	0.3733

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	549.6666667	15.81	0.0001	34.76623931
SEMANA	5.2000000	0.32	0.7483	16.09364158
IND1	-22.0000000	-0.45	0.6570	49.16688715
IND2	-6.6666667	-0.14	0.8928	49.16688715
SEMANA*IND1	33.3000000	1.46	0.1515	22.75984618
SEMANA*IND2	20.5000000	0.90	0.3733	22.75984618

Por lo dicho al comparar los modelos de los tres grupos, para determinar si los modelos difieren en ordenada al origen y/o en pendiente se plantea:

$$\mathrm{H}_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$
$$\mathrm{H}_1 : \text{al menos uno de los } \beta_i \neq 0 \ , \ i = 2,3,4,5$$

que postula en la hipótesis alternativa que las rectas para los grupos difieren en pendiente y/o en ordenada al origen. El estadístico para este test es

$$F = \frac{(SC_E(Mod.Red.) - SC_E(Mod.Compl.)) / (gl_E(Mod.Red.) - gl_E(Mod.Compl.))}{CM_E(Mod.Compl.)}$$

con distribución F con grados de libertad apropiados.

El Modelo Completo es el (1) ya ajustado con la salida 1). De ahí se obtiene $SC_E(Mod.Compl.) = 101012,067$ con 39 g.l. y $CM_E(Mod.Compl.) = 2590.053$. Si se modifica en el programa de SAS el comando MODEL sacando las interacciones y las variables indicadoras, se está ajustando el modelo

$$y = \beta_0 + \beta_1 x_1 + \epsilon \qquad (2)$$

que da por resultado:

Dependent Variable: PESO

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	16054.533	16054.533	5.61	0.0224
Error	43	123022.044	2860.978		
Corrected Total	44	139076.578			

R-Square	C.V.	Root MSE	Y Mean
0.115437	9.121783	53.488	586.38

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SEMANA	1	16054.533	16054.533	5.61	0.0224

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Model	1	16054.533	16054.533	5.61	0.0224

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	540.1111111	25.60	0.0001	21.09599459
SEMANA	23.1333333	2.37	0.0224	9.76554791

de donde se obtiene $SC_E(Mod.Red.) = 123022,044$ con 43 g.l.

Entonces,

$$F = \frac{(123022{,}044 - 101012{,}067)/(43-39)}{2590{,}053} = \frac{22009{,}977/4}{2590.053} = \frac{5502{,}49425}{2590.053} = 2.12$$

que al compararse con $F_{0.95;2;39} \approx 2.61$ lleva a No rechazar H_0 por lo cual

No puede afirmarse que las rectas correspondientes a los 3 grupos no coiciden en pendiente y/o en ordenada al origen.

Por lo tanto debemos ajustar el modelo (2)

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

el cual presenta al peso sólo explicado por el tiempo, es decir los pesos de los 3 grupos son ajustados por una única recta de regresión, que por la salida 2) es

$$\widehat{y} = 540,11111 + 23,13333 x_1$$

Para probar que el tiempo explica linealmente al peso planteamos

$$H_0 : \beta_1 = 0 \qquad\qquad H_1 : \beta_1 \neq 0$$

lo cual puede verse en la salida 2) en el renglón "SEMANA", del cual obtenemos un valor del estadístico $t = 2,37$ con $p = 0,0224$, lo que lleva a rechazar H_0 y concluir que *el tiempo explica linealmente al peso de todos los cerdos.*

Para probar si la recta pasa o no por el origen planteamos

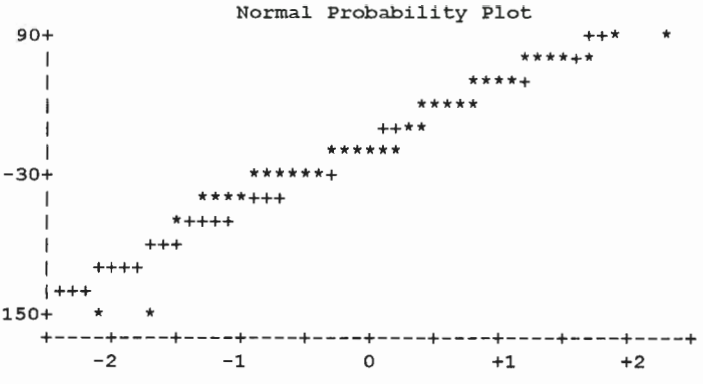
$$H_0 : \beta_0 = 0 \qquad\qquad H_1 : \beta_0 \neq 0$$

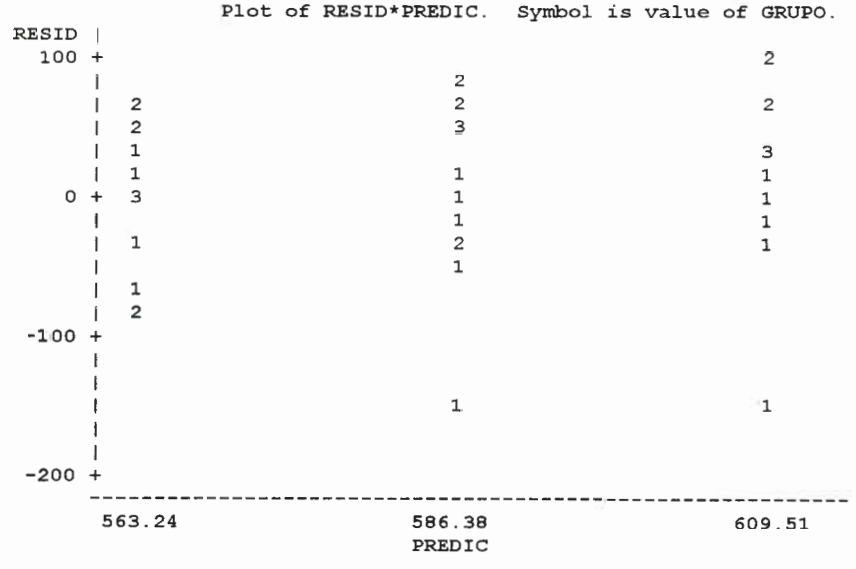
lo cual puede verse en la salida 2) en el renglón "INTERCEPT", del cual obtenemos un valor del estadístico $t = 25,60$ con $p = 0,0001$, lo que lleva a rechazar H_0 y concluir que *la recta no pasa por el origen.*

Luego el modelo (2) es el que deberíamos considerar.

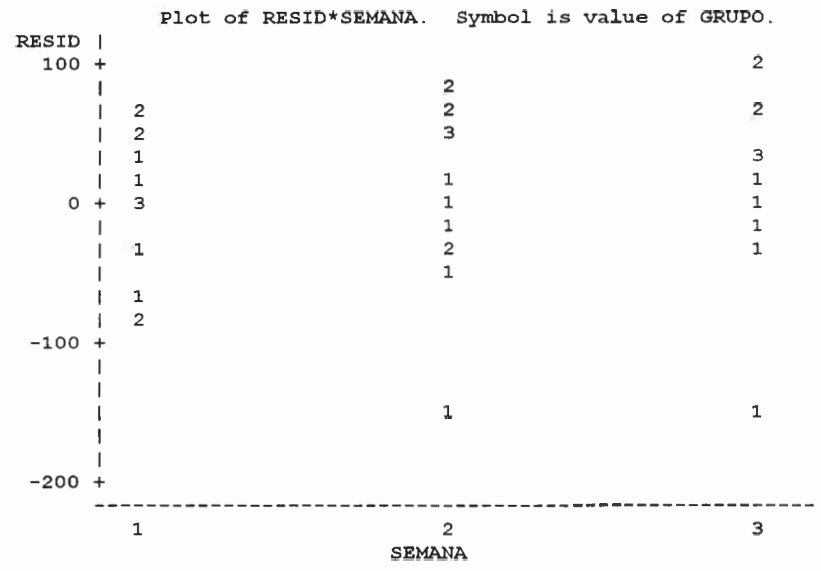
La adecuación al modelo la probamos en base a los residuos con:

```
                    Univariate Procedure

Variable=RESID

                          Moments

N                  45    Sum Wgts           45
Mean                0    Sum                 0
Std Dev       52.8768    Variance      2795.956
Skewness     -0.66512    Kurtosis      1.004698
USS            123022    CSS             123022
CV                  .    Std Mean      7.882407
T:Mean=0            0    Pr>|T|          1.0000
Num ^= 0           45    Num > 0            20
M(Sign)          -2.5    Pr>=|M|         0.5515
Sgn Rank         27.5    Pr>=|S|         0.7601
W:Normal     0.949354    Pr<W            0.0746
```





NOTE: 20 obs hidden.



NOTE: 20 obs hidden.

La normalidad de los residuos no se cumple, según muestra el test de normalidad (con un valor $p = 0,0746$), el gráfico de normalidad (con valores luy alejados de la recta), y los gráficos de Residuos versus Predichos y Residuos vs. Tiempo (que mustran que los residuos no son aleatorios y tienen valores muy alejados de cero).

Bibliografía

- Anderson, T.W. (1958) *Introduction to Multivariate Statistical Analysis*. John Wiley & Sons. New York.
- Crowder, M.J. y Hand, D.J. (1990) *Analysis of Repeted Measures*. Chapman&Hall. London.
- Dunn, O.J. y Clark, V.A. (1987) *Applied Statistics: Analysis of Variance and Regression*, 2nd. ed.. John Wiley & Sons. New York.
- Frison, L. y Pococh, S.J. (1992) Repeated measures in clinical trials: analysis using mean summary statistics and its implication for design. *Statistics in Medicine* **11**, 1685-704.
- Gorbein, J.A., Lazaro, C.G. y Little, R.J.A. (1992) Incomplete data in repeated measures analysis. *Statistical Methods in Medical Research* **1**, 275-295.
- Greenhouse, S.W. y Geisser, S. (1959) On the methods in the analysis of profile data. *Psychometrika* **24**, 95-112.
- Hand, D.J. y Crowder, M.J. (1996) *Practical Longitudinal Data Analysis*. Chapman&Hall. London.
- Hand, D.J. y Taylor, C.C. (1987) *Multivariate Analysis of Variance and Repeated Measures. A Practical Approach for Behavioural Scientists*. Chapman&Hall. London.
- Heyting, A., Tolboom, J.T.B.M. y Essers, J.G.A. (1992) Statistical handling of dropouts in longitudinal clinical trials. *Statistics in Medicine* **11**, 2043-62.
- Huynh, H. y Feldt, L.S. (1976) Estimation of the Box correction for degrees of freedom for sample data in randomised block and split-plot designs. *J. Educational Statist.* **1**, 69-82.
- Khatri, C. G. (1966) A note on a Manova model applied to problems in growth curve. *Ann. Inst. Statist. Math.* **18**, 75-86.
- Liang, K.Y. y Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.
- Mc Cullagh, P. (1983) Quasi-likelihood functions. *Ann. Statist.* **11**, 59-67.
- Mc Cullagh, P. y Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. Chapman&Hall. London.
- Mead, R., Curnow, R.N. y Hasted, A.M. (1993) *Statistical Methods in Agriculture and Experimental Biology*, 2nd. ed.. Chapman&Hall. London.
- Mendenhall, W. y Sincich, T. (1996) *A Second Course in Statistics: Regression Analysis*, 5th. ed..Prentice Hall, Upper Saddle River. New Jersey.

- Mendehall, W., Wackerly, D.D. y Scheaffer, R.L. (1994) *Estadística Matemática con Aplicaciones*, 2da. ed.. Grupo Editorial Iberoamérica. México.
- Montgomery, D.C. y Peck, E.A. (1982) *Introduction to Linear Regression Analysis*. John Wiley & Sons. New York.
- Rao, C.R. (1959) Some problems involving linear hypotheses in multivariate analysis. *Biometrika* **46**, 49-58.
- Snedecor, G.W. y Cochran, W.G. (1978) *Métodos Estadísticos*, 5a.Impresión en español, 1a.Ed.. Compañía Editorial Continental S.A.. México.
- Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439-47.
- Wilcox, C.J., Tatcher, W.W. y Martin, F.G. Statistical Analysis of Repeated Measurements in Physiology Experiments. *Journal of the Florida Agricultural Experiment Station* **9552**, 141.
- Winer, B.J., Brown, D.R. y Michels, K.M. (1991) *Statistical Principles in Experimental Design*, 3rd. ed.. McGraw-Hill. New York.
- Yandell, B.S. (1997) *Practical Data Analysis for Designed Experiments*, 1st. Ed.. Chapman&Hall. London.