UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Prueba para un conjunto de parámetros de un modelo de regresión: ejemplos de aplicación

Apellidos, nombre	Chirivella González, Vicente (vchirive@deio.upv.es)
Departamento	Estadística e Investigación Operativa Aplicadas y Calidad
Centro	Facultad de Administración y Dirección de Empresas

1 Resumen de las ideas clave

En este artículo docente trabajamos un enfoque que es complementario a la interpretación tradicional de los parámetros de un modelo de regresión, puesto que además del significado individual de cada parámetro, es posible que un cierto número de ellos tenga un significado conjunto. Si ese significado nos interesa, y hacemos afirmaciones sobre el mismo, necesitaremos realizar algún tipo de prueba de hipótesis para aceptar o rechazar dicha afirmación. Esta prueba es la prueba de hipótesis para un conjunto de parámetros. Aplicada convenientemente a nuestros datos, extraerán información de los mismos, información que utilizaremos para aceptar o rechazar la afirmación realizada.

2 Introducción

Los parámetros de un modelo de regresión cuantifican la relación existente entre la variable explicada y las variables explicativas que lo componen. Cada uno de ellos es, en esencia, el incremento de la variable explicada por incremento de la respectiva variable explicativa. Pero además de su significado individual, es posible que un cierto número de parámetros tenga un sentido conjunto. Pongamos que necesitamos un modelo que explique las ventas de tabletas SAMSUNG. Además de la renta de los consumidores y del precio de las propias tabletas, es posible incluir los precios de otras marcas: ACER, APPLE, MEDION, SONY, LENOVO,... Los parámetros que acompañan a estas otras marcas cuantificarán el efecto del precio de cada marca sobre las ventas de SAMSUNG. Pero si te fijas, todos estos "otros" parámetros tienen también un significado conjunto. ¿Y cuál es ese significado? Pues que son la competencia. De esta forma, y atendiendo a dichos parámetros, podríamos determinar si la competencia afecta, o no, a las ventas de nuestro producto.

Para averiguar si estamos afectados por la competencia tendremos que hacer una prueba de hipótesis, la prueba de un conjunto de parámetros. Así, escogiendo ciertos parámetros del modelo, y dándoles un valor adecuado (restricción lineal), determinaremos si efectivamente se acepta o no la afirmación que hayamos planteado. Si afirmamos que la competencia no influye en nuestras ventas, entonces el conjunto de parámetros β_{ci} (*ci*-competidor *i*) que multiplican a dichos precios deben ser cero ($\beta_{c1}= \beta_{c2}=...=0$). Y es que, al fin y al cabo, si esos parámetros valen cero, ¡es que no importan los precios de los competidores!

3 Objetivos

Los objetivos marcados para este artículo docente son:

1. Utilizaremos el significado de los parámetros del modelo para encontrar puntos en común entre ellos, y los interpretaremos en el ámbito del problema analizado.
2. Plantearemos y realizaremos la prueba de un conjunto de parámetros para aceptar o rechazar ciertas afirmaciones que hagamos sobre ese significado común de los parámetros. Prestaremos especial atención a las restricciones que debemos proponer, a los dos modelos que debemos estimar, y al uso de los resultados intermedios que hemos obtenido para realizar finalmente la prueba.

4 Prueba para un conjunto de parámetros

En este apartado vamos presentar la prueba para un conjunto de parámetros y la aplicaremos a ejemplos concretos. Analizaremos dos modelos, uno que incluye variables cualitativas y otro que no lo hace, y veremos cómo los significados individuales de sus parámetros muestran posibles significados conjuntos. Tras este análisis sacaremos algunas conclusiones sobre lo que hemos hecho y las generalizaremos. Finalmente propondremos nuevos modelos para ejercitar los conocimientos adquiridos.

4.1 Primer ejemplo: Cantidad de vidrio reciclado en la Comunidad Valenciana

Comenzamos presentando el ejemplo que utilizaremos para introducir los conceptos:

Se desea proponer un modelo que explique la cantidad de vidrio reciclado en las tres provincias de la Comunidad Valenciana a partir del número de contenedores instalados para su recogida. La cantidad de vidrio reciclado, PTVIDRIO, se mide en Kg, mientras que la cantidad de contenedores, NCONT, son unidades físicas. Para cuantificar las diferencias entre provincias se define la variable ficticia CAS, que toma el valor 1 si el dato corresponde a Castellón, y cero en otro caso, y la variable ficticia VAL, que toma el valor 1 si el dato corresponde a Valencia y 0 en otro caso, quedando como referencia la provincia de Alicante. Finalmente se opta por restar 1000 al número de contenedores, con lo que el modelo final es:

$$PTVIDRIO = \beta_0 + \beta_1 CAS + \beta_2 VAL + \beta_3 (NCONT - 1000)$$
$$+\beta_4 CAS * (NCONT - 1000) + \beta_5 VAL * (NCONT - 1000) + U$$

Las cuestiones que se plantean son:

a) ¿La provincia afecta a la relación entre la cantidad de vidrio recogido y el número de contenedores?

b) Cuando hay 1000 contenedores en la calle, ¿la cantidad de vidrio recogido (en promedio) es la misma en las tres provincias?

c) ¿Se recoge la misma cantidad de vidrio (en promedio) por cada contenedor que se instala en la calle en las tres provincias?

d) Por cada nuevo contenedor en la calle, ¿se recoge en Valencia la misma cantidad de vidrio (en promedio) que en Alicante y Castellón juntas?

El primer paso para responder a las preguntas será determinar el significado de los parámetros del modelo, pues las preguntas tienen que ver con uno o más de ellos:

β_0 promedio de la cantidad de vidrio recogido en Alicante cuando hay 1000 contenedores para hacerlo.

β_3 incremento de la cantidad promedio de vidrio recogido en Alicante por cada contenedor instalado.

β_1 diferencia de la cantidad promedio de vidrio recogido en Castellón respecto a Alicante cuando hay 1000 contenedores para hacerlo.

β_4 diferencia del incremento de la cantidad promedio de vidrio recogido por cada contenedor instalado en Castellón respecto a Alicante.

β_2 diferencia de la cantidad promedio de vidrio recogido en Valencia respecto a Alicante cuando hay 1000 contenedores para hacerlo.

β_5 diferencia del incremento de la cantidad promedio de vidrio recogido por cada contenedor instalado en Valencia respecto a Alicante.

Entonces:

a) ¿La provincia afecta a la relación entre la cantidad de vidrio y el número de contenedores?

En la relación entre la cantidad de vidrio recogido y el número de contenedores, hay ciertos parámetros que cuantifican diferencias entre provincias. Estos parámetros son, de acuerdo a su definición, β_1, β_2, β_4 y β_5, pues podemos ver que todos ellos comienzan por la palabra "diferencia". ¿Qué ocurriría si todos ellos valiesen cero a la vez? Pues que las diferencias serían cero, esto es, que no existirían diferencias entre las provincias, al menos en el sentido que dicen los parámetros. Por lo tanto, si β_1=0, β_2=0, β_4=0 y β_5=0 entonces no habrían diferencias entre provincias, y éstas no afectarían a la relación entre cantidad recogida y número de contenedores.

b) Cuando hay 1000 contenedores en la calle, ¿la cantidad de vidrio recogido es la misma en las tres provincias?

Los parámetros que incluyen en su definición el hecho de haber 1000 contenedores en la calle son β_0, β_1 y β_2, y de éstos, los que cuantifican diferencias entre provincias son β_1 y β_2, por incluir además la palabra "diferencia" en su definición. Si estos dos parámetros fuesen cero a la vez entonces, ¿querría eso decir que no existirían diferencias entre las provincias? Pues si, ¿no? Por lo tanto, si β_1=0 y β_2=0 entonces no existen diferencias entre provincias (cantidad de vidrio) cuando éstas tienen 1000 contenedores en la calle.

c) ¿Se recoge la misma cantidad de vidrio (en promedio) por cada contenedor que se instala en la calle en las tres provincias?

Ahora buscamos en las definiciones de los parámetros el hecho de que se esté cuantificando la cantidad de vidrio recogida por cada contenedor. Estos parámetros son β_3, β_4 y β_5, y de éstos, los que cuantifican diferencias entre provincias son β_4 y β_5, por incluir además la palabra "diferencia" en su definición. Al igual que antes, si estos dos parámetros valiesen cero a la vez entonces esto querría decir que no existirían diferencias entre las provincias. Por lo tanto, si β_4=0 y β_5=0 entonces no existen diferencias entre provincias en cuanto a la cantidad de vidrio recogida por contenedor.

d) Por cada nuevo contenedor en la calle, ¿se recoge en Valencia la misma cantidad de vidrio que Alicante y Castellón juntas?

En este apartado seguimos cuantificando la cantidad de vidrio recogida por cada contenedor, por lo que los parámetros involucrados siguen siendo β_3, β_4 y β_5. De acuerdo al modelo, y a sus definiciones, la cantidad de vidrio recogida por contenedor en Alicante es β_3, en Castellón es $\beta_3+\beta_4$ y en Valencia es $\beta_3+\beta_5$. De esta forma, la cantidad de Alicante y Castellón juntas es $\beta_3+(\beta_3+\beta_4)$ o $2\beta_3+\beta_4$, y según se plantea en la pregunta, deseamos comprobar si esta cantidad, $2\beta_3+\beta_4$ (Alicante y Castellón), es igual a $\beta_3+\beta_5$ (Valencia), así que $2\beta_3+\beta_4=\beta_3+\beta_5$ o bien que $\beta_3+\beta_4-\beta_5=0$. Si se cumple esta condición, la cantidad de vidrio recogida por contenedor en Valencia es la misma que la que se recoge conjuntamente en Alicante y Castellón.

Bueno, pues ahora que sabemos qué parámetros, y qué valores deben tomar, habrá que realizar la prueba de hipótesis...

4.2 Prueba de hipótesis para un conjunto de parámetros

La prueba de hipótesis sobre un conjunto de parámetros se basa en un estadístico *F*. Para realizar la prueba debemos hacer el ajuste de dos modelos. En primer lugar, ajustaremos el modelo sin las restricciones, el modelo original (por así decirlo), y anotaremos su suma de cuadrados residual y la llamaremos *SCRc*, y anotaremos también los grados de libertad del error. En segundo lugar, ajustaremos el modelo con las *s* restricciones que debemos comprobar, y anotamos una segunda suma de cuadrados residual, que llamaremos *SCRr*. En la siguiente expresión, se supone que hay *s* parámetros que deberían valer cero, pero se puede cambiar la hipótesis nula para que sea otra condición. La prueba se basa en el cálculo de un estadístico F_{calc}, efectuado de la siguiente manera:

H_0 $\beta_1 = \beta_2 = \ldots = \beta_s = 0$

H_1 al menos un parámetro es distinto de cero

$$F_{calc} = \frac{(SCR_r - SCR_c)/s}{SCRc/(n-k-1)} = \frac{\Delta SCR/s}{SCRc/(n-k-1)} \equiv F_{s\ n-k-1}$$

Si $F_{calc} \leq F_{s,n-k-1}^{\alpha}$ entonces se acepta H_0 ($\beta_1 = \beta_2 = \ldots = \beta_s = 0$)

Ahora aplicaremos todo esto a nuestro problema. Para no hacerlo demasiado largo, no contestaremos a todas las preguntas planteadas.

a) ¿La provincia afecta a la relación entre la cantidad de vidrio y el número de contenedores?

Para realizar la prueba indicamos en primer lugar la hipótesis nula y la alternativa. La hipótesis nula recoge la opción de "todo es igual", "nada cambia", "no hay influencia", de forma que esperamos que los datos permitan rechazarla y afirmar lo contrario.

H_0 $\beta_1=\beta_2=\beta_4=\beta_5=0$	**La provincia no afecta** a la relación entre la cantidad de vidrio recogida y el número de contenedores.
H_1 Al menos uno de los parámetros es distinto de cero	**La provincia si afecta** a la relación entre la cantidad de vidrio recogida y el número de contenedores.

El modelo original es

$$PTVIDRIO = \beta_0 + \beta_1 CAS + \beta_2 VAL + \beta_3(NCONT - 1000)$$
$$+\beta_4 CAS * (NCONT - 1000) + \beta_5 VAL * (NCONT - 1000) + U$$

Una vez realizado el ajuste, la suma de cuadrados residual es de *SCRc*=7,8785138E13, y los grados de libertad residuales, 39 (no presentamos el ajuste en el documento).

El modelo restringido debe contener las afirmaciones que hemos hecho, que $\beta_1=\beta_2=\beta_4=\beta_5=0$, por lo que debemos sustituir estos valores en el modelo, lo cual implica eliminar sus correspondientes variables. Hemos pedido que cuatro parámetros valgan cero, por lo que tenemos 4 restricciones (*s*):

$$PTVIDRIO = \beta_0 + \cancel{\beta_1 CAS + \beta_2 VAL} + \beta_3(NCONT - 1000)$$
$$\cancel{+\beta_4 CAS * (NCONT - 1000) + \beta_5 VAL * (NCONT - 1000)} + U$$

con lo que queda

$$PTVIDRIO = \beta_0 + \beta_3(NCONT - 1000) + U$$

Realizado el ajuste del modelo, la suma de cuadrados residual es de *SCRr*= 2,5602261E14.

Por tanto:

$$F_{calc} = \frac{(SCR_r - SCR_c)/s}{SCRc/(n-k-1)} = \frac{(2{,}5602261E14 - 7{,}8785138E13)/4}{7{,}8785138E13/39} = 21{,}9$$

Como $F_{calc} = 21{,}9 > F_{4,39}^{0,05} = 2{,}61$ entonces se rechaza H_0 y por tanto **la provincia si afecta a la relación entre la cantidad de vidrio recogida y el número de contenedores**. Otra cuestión será explicar la forma en que afecta. Pero afectar, afecta.

b) Cuando hay 1000 contenedores en la calle, ¿la cantidad de vidrio recogido es la misma en las tres provincias?

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

H_0 $\beta_1=\beta_2=0$	La cantidad de vidrio recogido cuando hay 1000 contenedores **es la misma** en las tres provincias.
H_1 Al menos uno de los parámetros es distinto de cero	La cantidad de vidrio recogido cuando hay 1000 contenedores **no es la misma** en las tres provincias.

No hay cambios en el modelo original, así que nos sirven los valores anteriores de suma de cuadrados y grados de libertad residuales. El modelo restringido ahora debe contener las afirmaciones que hemos hecho, que $\beta_1=\beta_2=0$, por lo que debemos sustituirlas en el mismo, lo cual implica eliminar sus correspondientes variables. Hemos pedido que dos parámetros valgan cero, por lo que tenemos dos restricciones.

$$PTVIDRIO = \beta_0 + \cancel{\beta_1 CAS + \beta_2 VAL} + \beta_3(NCONT - 1000)$$
$$+\beta_4 CAS * (NCONT - 1000) + \beta_5 VAL * (NCONT - 1000) + U$$

con lo que queda

$$PTVIDRIO = \beta_0 + \beta_3(NCONT - 1000) + \beta_4 CAS * (NCONT - 1000)$$
$$+\beta_5 VAL * (NCONT - 1000) + U$$

Realizado el ajuste del modelo, la suma de cuadrados residual es de $SCRr$=1,2207093E14.

Por tanto:

$$F_{calc} = \frac{(SCR_r - SCR_c)/s}{SCRc/(n-k-1)} = \frac{(1{,}2207093E14 - 7{,}8785138E13)/2}{7{,}8785138E13/39} = 10{,}7136$$

Como $F_{calc} = 10{,}7136 > F_{2,39}^{0,05} = 3{,}2381$ entonces se rechaza H_0 y por tanto **la cantidad de vidrio recogido cuando hay 1000 contenedores es diferente en las provincias**.

Para determinar las diferentes cantidades recogidas, necesitamos los valores de las estimaciones de β_0, β_1 y β_2. En el correspondiente ajuste (que no presentamos) se tiene que la estimación para β_0 es de 7070839,2 Kg, la de β_1 es de -4731455,7 Kg y la de β_2 es de -5878048,2 Kg. Por lo tanto, cuando hay 1000 contenedores en la calle, en Alicante se recogen 7070839,2 Kg, en Castellón se recoge algo menos que en Alicante, 7070839,2-4731455,7=2339383,5 Kg, y en Valencia también menos que en Alicante, 7070839,2-5878048,2=1192791 Kg.

Aunque los números de Castellón y Valencia parecen diferentes, en realidad no lo sabemos con certeza, dado que estos valores son muestrales y no poblacionales. Es necesario realizar una prueba de hipótesis adicional para saber si las dos provincias son realmente diferentes. Necesitamos de una prueba adicional porque los parámetros del modelo cuantifican diferencias entre Alicante y Castellón, y entre Alicante y Valencia, pero no entre Castellón y Valencia, así que no hay ningún parámetro que en su definición relacione a Castellón y Valencia.

La cantidad de vidrio recogido en Castellón cuando hay 1000 contenedores es, como acabamos de ver, $\beta_0+\beta_1$, mientras que la cantidad de vidrio recogida en Valencia en las mismas circunstancias es $\beta_0+\beta_2$. Así que si nos preguntamos si ambas cantidades son iguales entonces debe ocurrir que $\beta_0+\beta_1=\beta_0+\beta_2$, o simplificando, que $\beta_1=\beta_2$, o finalmente que $\beta_2-\beta_1=0$. Para responder a esta pregunta seguimos teniendo la prueba de un conjunto de parámetros, sólo que ahora la hipótesis nula es $\beta_2-\beta_1=0$ y significa que las cantidades de vidrio recogidas en Castellón y Valencia son las mismas.

H_0 $\beta_2-\beta_1=0$ $\beta_2=\beta_1$	La cantidad de vidrio recogida en Castellón y Valencia **es la misma** cuando hay 1000 contenedores disponibles.
H_1 $\beta_2-\beta_1\neq 0$ $\beta_2\neq\beta_1$	La cantidad de vidrio recogida en Castellón y Valencia **no es la misma** cuando hay 1000 contenedores disponibles.

Y el nuevo modelo restringido, al sustituir la restricción propuesta:

$$PTVIDRIO = \beta_0 + \beta_2 CAS + \beta_2 VAL + \beta_3(NCONT - 1000)$$
$$+\beta_4 CAS * (NCONT - 1000) + \beta_5 VAL * (NCONT - 1000) + U$$

y agrupando

$$PTVIDRIO = \beta_0 + \beta_2(CAS + VAL) + \beta_3(NCONT - 1000)$$
$$+\beta_4 CAS * (NCONT - 1000) + \beta_5 VAL * (NCONT - 1000) + U$$

Cuando se realice el ajuste del nuevo modelo, que incluye una única restricción, se dispondrá de la nueva *SCRr* y con ello la prueba de hipótesis que permitirá aceptar o rechazar la nueva afirmación.

Procederíamos de la misma forma con las otras dos cuestiones. Aunque no facilitamos las nuevas sumas de cuadrados residuales, al menos puedes plantear las hipótesis y obtener el modelo restringido. ¿Cuál sería el modelo restringido del apartado d)?

4.3 Segundo ejemplo: el "efecto carajillo" en el consumo de café

Planteamos un nuevo modelo, ahora sin variables cualitativas:

El siguiente modelo explica el consumo de café a través de la renta de los consumidores, del precio del café, de los precios del azúcar, de la leche, del anís, del coñac y del ron. Plantea una prueba de hipótesis que permita aceptar que los precios de las bebidas alcohólicas no influyen en el consumo de café.

$$C.Café = \beta_0 + \beta_1 Renta + \beta_2 P.Café + \beta_3 P.Azúcar + \beta_4 P.Leche$$

$$+\beta_5 P.Anís + \beta_6 P.Coñac + \beta_7 P.Ron + U$$

Como la pregunta se refiere a las bebidas alcohólicas, definiremos únicamente los parámetros del modelo que acompañan a las bebidas alcohólicas. De forma general, estos parámetros se definen como:

β_5 β_6 β_7 incremento de la cantidad promedio de café consumido por cada euro que aumenta el precio del #venenofavorito cuando el resto de variables se mantienen constantes.

donde #venenofavorito representa a cada una de las bebidas alcohólicas, y todos estos parámetros se interpretan a renta y a precios (restantes) constantes.

Si hacemos que los parámetros β_5, β_6 y β_7 sean cero a la vez, eso significaría que los precios de las bebidas alcohólicas no influyen en el consumo de café. Bastaría que uno de estos parámetros fuera distinto de cero, para afirmar que si que lo hacen.

H_0 $\beta_5=\beta_6=\beta_7=0$	El precio de las bebidas alcohólicas **no influye** en el consumo de café.
H_1 Al menos uno de los parámetros es distinto de cero	El precio de las bebidas alcohólicas **si influye** en el consumo de café.

Aunque en caso de rechazar la hipótesis nula no sabremos cuál de los "venenos favoritos" estudiados influye en el consumo de café.

4.4 El "arte" de plantear las restricciones de parámetros

Una vez analizado los ejemplos de la cantidad de vidrio recogido para reciclar y del consumo de café, cabría sacar algunas conclusiones de ellos:

1.- Es imprescindible conocer el significado de los parámetros del modelo, pues la pregunta que planteemos estará relacionada, de alguna forma, con el significado de los parámetros.

2.- A partir de los significados de los parámetros podemos plantear la pregunta que deseemos, aunque en realidad deberíamos interpretar esto al revés, es decir... a partir de la pregunta nos planteamos, ¿cuál es el modelo y los parámetros que debemos analizar?

3.- Las definiciones de los parámetros son las que nos permiten encontrar afinidades entre ellos, ya que algunos:

   a. significan lo mismo, salvo que se refieren a una variable explicativa diferente, variables que pueden tener algo en común

   b. cuantifican una diferencia respecto al grupo escogido como referencia

4.- Cuando en el modelo pretendemos comparar entre sí tres o más grupos, será necesario plantear nuevas restricciones a los parámetros, puesto que éstos no comparan entre sí a grupos diferentes del grupo de referencia.

4.5 Más ejemplos de aplicación

Planteamos ahora unos modelos, y realizamos unas preguntas sobre ellos, para que selecciones los parámetros y las relaciones que deben existir para contestarlas. Cuando definas los parámetros podrías hacer uso de un comodín como en el ejemplo #venenofavorito para evitar tener que definirlos todos con detalle. En el siguiente apartado tendrás las respuestas.

P1.-Los principales productores de aceite de oliva de la CEE, en orden descendente, España, Italia y Grecia. En los últimos años España se consolida como la principal potencia de producción y exportación de aceite de oliva, y en particular en el año 2013 España triplica la producción de Italia. Para cuantificar las diferencias entre países a este respecto, se ha propuesto y ajustado un modelo que relaciona la producción nacional con la producción total en la Comunidad Europea. Así la variable PRODUCCIÓN es la producción de cada uno los tres países (en toneladas métricas), TOTAL es la producción total de la Comunidad Europea (en toneladas métricas), y se crean las variables ficticias correspondientes a los tres países para ser introducidas adecuadamente en el modelo. También se ha restado a la variable TOTAL la producción total de la comunidad del año 2013, quedando como TOTAL13.

$$PRODUCCION = \beta_0 + \beta_1 ESPAÑA + \beta_2 GRECIA + \beta_3 TOTAL13$$
$$+\beta_4 ESPAÑA * TOTAL13 + \beta_5 GRECIA * TOTAL13 + U$$

Plantear las pruebas de hipótesis que permitan determinar:

a) Si es aceptable la afirmación realizada de que España triplica la producción de Italia en el año 2013.

b) El porcentaje de producción de aceite de oliva que corresponde a los tres países es el mismo.

P2.-En el informe socioeconómico del sector de la cerveza en España de 2010, publicado por la Asociación de Cerveceros de España, se detalla el destino geográfico de la cerveza producida, y el porcentaje que supone cada una de las siete zonas existentes en el total de la producción. Para comparar las cantidades de cerveza producidas por cada zona se propone el siguiente modelo:

$$PRODZONA = \beta_0 + \beta_1 ZONA1 + \beta_2 ZONA2 + \beta_3 ZONA3 + \beta_4 ZONA5 + \beta_5 ZONA6 + \beta_6 ZONA7 \\ + \beta_7 (PRODTOTAL - 2\ 10^6) + \beta_8 ZONA1 * (PRODTOTAL - 2\ 10^6) + \beta_9 ZONA2 \\ * (PRODTOTAL - 2\ 10^6) + \beta_{10} ZONA3 * (PRODTOTAL - 2\ 10^6) + \beta_{11} ZONA5 \\ * (PRODTOTAL - 2\ 10^6) + \beta_{12} ZONA6 * (PRODTOTAL - 2\ 10^6) + \beta_{13} ZONA7 \\ * (PRODTOTAL - 2\ 10^6) + U$$

donde PRODZONA es la producción de cerveza según la zona, y PRODTOTAL es la producción total de cerveza, ambas medidas en hectolitros. Se crean las variables ficticias necesarias para introducir la zona en el modelo:

ZONA	
1	*Noreste e I. Baleares*
2	*C. Valenciana, Albacete y Murcia*
3	*Andalucía, sur Extremadura, Ceuta y Melilla*
4	*Centro*
5	*Noroeste*
6	*Norte, y norte de Castilla y León*
7	*I. Canarias*

Plantear la prueba de hipótesis que permita determinar:

a) si la cuota de mercado de las zonas 1, 2, 3 y 4 son iguales entre sí.

b) si la cuota de mercado conjunta de las zonas 5, 6 y 7 es igual a la de la zona 1.

c) si cuando se producen $2 \cdot 10^6$ hectolitros en total, la cantidad de cerveza producida en conjunto por las zonas bañadas por el Mediterráneo es la misma que la del resto de zonas.

d) si cuando se producen $2 \cdot 10^6$ hectolitros, las cantidades producidas en promedio en cada zona suman esos $2 \cdot 10^6$ hectolitros.

P3.-Se desea explicar el gasto realizado en unos determinados juegos de azar (GASTOTIPO, euros por habitante) en función del gasto total realizado en los mismos (GASTO500, gasto en total menos 500 euros). Los juegos de azar en España, según el INE, son: Primitiva, BonoLoto, Euromillones, Gordo de la Primitiva, Lotería Nacional, Quiniela, Bingo, Máquinas, Casino, ONCE, otros... Es posible crear variables que indiquen cada uno de los juegos, y utilizarlas en un modelo como el siguiente (la referencia escogida es otros juegos).

$$GASTOTIPO = \beta_0 + \beta_1 PRIMITIVA + \beta_2 BONOLOTO + \beta_3 EUROMILLONES + \beta_4 GORDO \\ + \beta_5 LOTERÍA + \beta_6 QUINIELA + \beta_7 MÁQUINAS + \beta_8 BINGO + \beta_9 CASINO \\ + \beta_{10} ONCE + \beta_{11} GASTO500 + \beta_{12} PRIMITIVA * GASTO500 \\ + \beta_{13} BONOLOTO * GASTO500 + \beta_{14} EUROMILLONES * GASTO500 \\ + \beta_{15} GORDO * GASTO500 + \beta_{16} LOTERÍA * GASTO500 + \beta_{17} QUINIELA \\ * GASTO500 + \beta_{18} MÁQUINAS * GASTO500 + \beta_{19} BINGO * GASTO500 \\ + \beta_{20} CASINO * GASTO500 + \beta_{21} ONCE * GASTO500 + U$$

Cuando el gasto total es de 500 euros por habitante,

a) ¿el gasto en juegos de titularidad pública (LAE) es el mismo que el gasto en juegos de titularidad privada?

b) ¿el gasto acumulado en juegos donde el resultado es debido a la intervención de bolas es el mismo que en el que no intervienen?

4.6 Resultados a las cuestiones planteadas

A continuación presentamos la hipótesis nula de los problemas planteados. Las expresiones no han sido simplificadas para poder ver mejor como se han obtenido, excepto la cuestión d) del P2, para contrastar la diferencia de hacerlo o no hacerlo.

P1.- Exportación de aceite de oliva

a) $\beta_0 + \beta_1 = 3\beta_0$
b) $\beta_4 = \beta_5 = 0$

P2.- Producción de cerveza

a) $\beta_8 = \beta_9 = \beta_{10} = 0$
b) $(\beta_7 + \beta_{11}) + (\beta_7 + \beta_{12}) + (\beta_7 + \beta_{13}) = (\beta_7 + \beta_8)$
c) $(\beta_0 + \beta_1) + (\beta_0 + \beta_2) + (\beta_0 + \beta_3) = \beta_0 + (\beta_0 + \beta_4) + (\beta_0 + \beta_5) + (\beta_0 + \beta_6)$
d) $7\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 = 2\ 10^6$

P3.- Gasto en juegos de azar

a) $(\beta_0 + \beta_1) + (\beta_0 + \beta_2) + (\beta_0 + \beta_3) + (\beta_0 + \beta_4) + (\beta_0 + \beta_5) + (\beta_0 + \beta_6) = (\beta_0 + \beta_7) + (\beta_0 + \beta_8) + (\beta_0 + \beta_9) + (\beta_0 + \beta_{10})$
b) $(\beta_0 + \beta_1) + (\beta_0 + \beta_2) + (\beta_0 + \beta_3) + (\beta_0 + \beta_4) + (\beta_0 + \beta_5) + (\beta_0 + \beta_6) + (\beta_0 + \beta_8) + (\beta_0 + \beta_{10}) = (\beta_0 + \beta_7) + (\beta_0 + \beta_9)$

5 Resumen y Conclusiones

- En este artículo hemos visto que, en ocasiones, los parámetros de un modelo de regresión tienen un significado conjunto además de su significado individual, que podemos hacernos preguntas sobre ese significado conjunto, y que para responderlas será necesario utilizar la prueba de un conjunto de parámetros.
- Aunque en los ejemplos planteados ocurría lo contrario, lo primero es la pregunta a responder, y luego viene encontrar el modelo y los parámetros del mismo que son necesarios para responder. Si no somos capaces de encontrar un modelo y unos parámetros que nos ayuden, no podremos responder a la pregunta.
- Es imprescindible conocer el significado de los parámetros del modelo, pues la pregunta que se plantee estará relacionada, de alguna forma, con su significado.
- Las definiciones de los parámetros son las que nos permiten encontrar afinidades entre ellos, ya que algunos:
    - significan lo mismo, salvo que se refieren a una variable explicativa diferente, variables que pueden tener algo en común
    - cuantifican una diferencia respecto al grupo escogido como referencia
- Cuando en el modelo pretendemos comparar entre sí tres o más grupos, será necesario plantear nuevas restricciones a los parámetros, puesto que éstos no comparan entre sí a grupos diferentes del grupo de referencia.

6 Bibliografía

D. Gujarati : "Basic Econometrics", Ed. McGrawHill - 4ª edition, páginas 297-334.

D. Peña: "Estadística: Modelos y Métodos. (Vol.2) Modelos lineales y Series temporales", Ed. Alianza Universidad-Textos, páginas 307-548.