Instituto Valenciano de Investigaciones Agrarias

Seminario

MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN AGRONÓMICA

Tema 9

MODELOS DE REGRESIÓN (II)

INCLUSIÓN DE VARIABLES CUALITATIVAS

MODELOS DE REGRESIÓN LINEAL (II)
INCLUSIÓN DE VARIABLES CUALITATIVAS

1. Inclusión de variables explicativas de tipo cualitativo
   - 1.1 Utilización de variables binarias
   - 1.2 Interpretación de los parámetros
2. Inclusión de interacciones con variables cualitativas
3. Ejemplo: Uso de un modelo de regresión en la comparación de dos poblaciones
4. Un test general sobre los parámetros de un modelo de regresión
5. Ejemplo: Efecto del día de la semana en el consumo de energía
6. Relación entre Modelos de Regresión y Análisis de la Varianza
7. Anova en modelos desequilibrados: medias “least squares”

Inclusión de variables explicativas de tipo cualitativo

- Tiene un **gran interés práctico**, la posibilidad de incluir variables explicativas de naturaleza cualitativa en modelos de regresión, puesto que ello potencia enormemente la posibilidad de modelar mediante los mismos fenómenos reales.
- Por ejemplo, en un estudio se han recogido datos sobre el rendimiento de cierto cultivo en un conjunto de explotaciones, en cada una de las cuales se cultivaba una de tres posibles variedades (codificadas como 1, 2 ó3), así como sobre la dosis empleada de abono
  - Y : Rendimiento
  - X1 : DOSIS de abono
  - X2 : VARIEDAD (codificada como1, 2 ó 3)
- Lo que i**NUNCA!** puede hacerse es formular un modelo incluyendo los códigos de una variable cualitativa (con más de dos posibles "valores") como si se tratara de una variable cuantitativa
- $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ **INCORRECTO**
- (Comprobar que en el modelo anterior los parámetros β_0 y $\beta_1$carecen de sentido)

¿Cómo se incluyen variables cualitativas en el modelo?

- La solución consiste en **crear, por cada variable cualitativa con K alternativas, K-1 nuevas variables, con únicos valores posibles 0 ó 1**, que indiquen en cada caso cuál es la alternativa correspondiente. Estas variables binarias se denominan en inglés variables "dummy", y a veces en castellano variables "ficticias".
- Por ejemplo, en el caso considerado, para incluir la variedad en el modelo se definirían las dos variables siguientes:
  - Z_2: vale 1 si la Variedad es la 2 y 0 en caso contrario
  - Z_3: vale 1 si la Variedad es la 3 y 0 en caso contrario

Variedad	Variables Dummy		
	Z_2	Z_3	
Variedad. 1	0	0	Valores de las variables dummy según la variedad
Variedad 2	1	0	
Variedad 3	0	1	

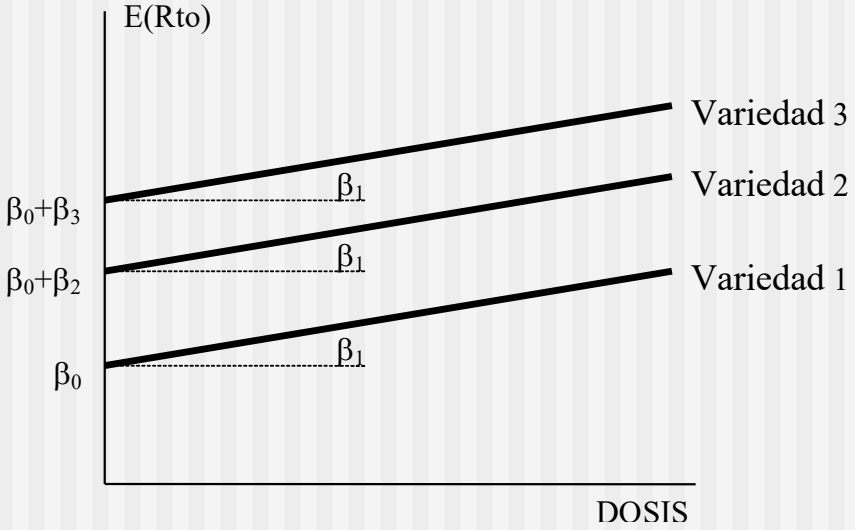
- <u>Nota</u>: Obsérvese que el número de variables "dummy" a crear, y por tanto el número de parámetros a introducir en el modelo, para considerar el efecto del factor cualitativo, coincide con los grados de libertad asociados a ese factor en un Anova

Interpretación de los parámetros del modelo

- El modelo: $\mathbf{E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 Z_2 + \beta_3 Z_3}$
- sí que es correcto, puesto que sus parámetros tienen un significado técnico concreto, que puede deducirse particularizando la ecuación anterior para cada variedad
  - E(Y/X1, Variedad 1) = $\beta_0 + \beta_1 X_1$
  - E(Y/X1, Variedad 2) = $\beta_0 + \beta_1 X_1 + \beta_2 = (\beta_0 + \beta_2) + + \beta_1 X_1$
  - E(Y/X1, Variedad 3) = $\beta_0 + \beta_1 X_1 + \beta_3 = (\beta_0 + \beta_3) + + \beta_1 X_1$
- Restando la primera ecuación de la segunda y de la tercera se deduce de forma inmediata:
  - β_2: diferencia del Rto medio obtenida cuando se cultiva la Variedad 2 respecto a cuando se cultiva la Variedad 1 (sea cual sea el valor X_1 de la DOSIS)
  - β_3: diferencia del Rto medio obtenida cuando se cultiva la Variedad 3 respecto a cuando se cultiva la Variedad 1 (sea cual sea el valor X_1 de la DOSIS)
- La hipótesis de que las tres variedades tienen el mismo rendimiento medio sería, por tanto, equivalente a: $\beta_2 = \beta_3 = 0$

Interpretación de los parámetros del modelo (continuación)

- La interpretación de los restantes parámetros del modelo (1) será:
  - β_0: Rto medio con la variedad 1 y sin abonar (o sea con $X_1=0$)
  - β_1 : incremento en E(Rto) por cada unidad que aumenta la dosis de abonado. El modelo asume que dicho incremento es idéntico en las tres variedades, o sea que el efecto del abonado no depende de la variedad (ausencia de interacción)
- En definitiva, el modelo (1) corresponde a una situación como la representada en la siguiente figura, en la que se ha asumido $\beta_3 > \beta_2$ y positivos ambos



Inclusión de posibles interacciones

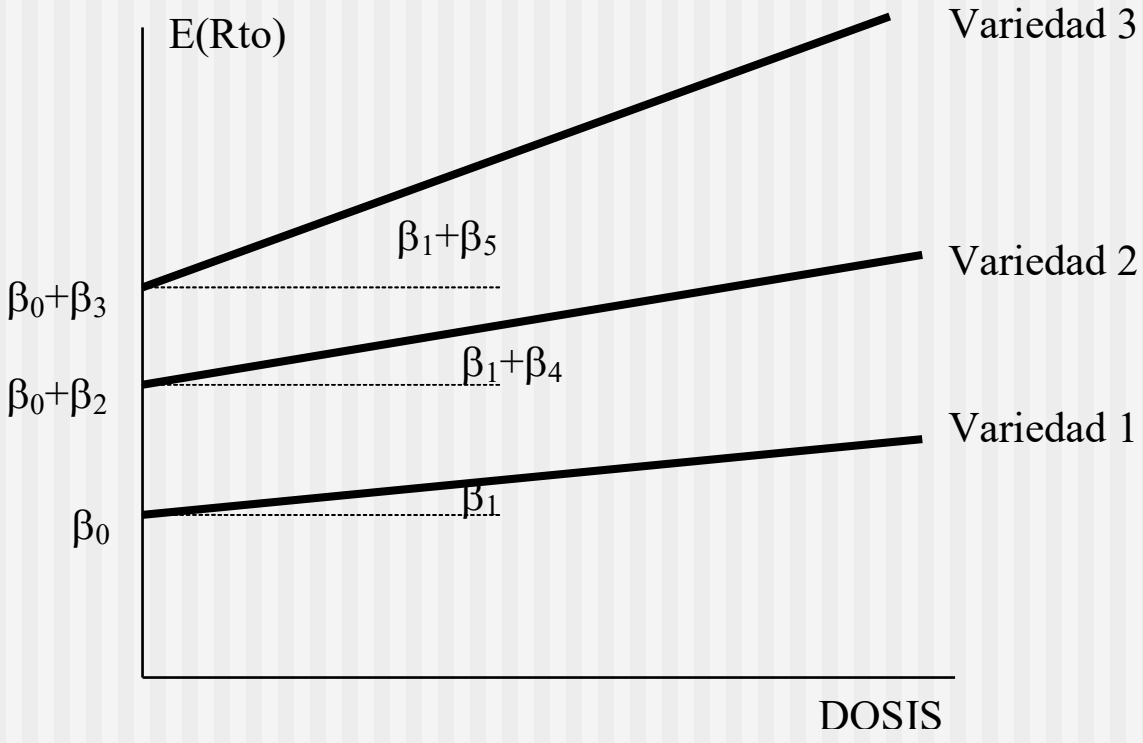
- Es posible formular un modelo más general que permita estudiar posibles interacciones entre las dos variables. Para ello basta incluir en el modelo nuevas variables que sean el producto de X_1 por cada una de las variables Z_k asociadas al efecto de la variedad

- En efecto el modelo: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 X_1 Z_2 + \beta_5 X_1 Z_3$ implica:

  - $E(Y/\text{Variedad } 1) = \beta_0 + \beta_1 X_1 + 0 + 0 + 0 + 0 = \beta_0 + \beta_1 X_1$
  - $E(Y/\text{Variedad } 2) = \beta_0 + \beta_1 X_1 + \beta_2 + 0 + \beta_4 X_1 + 0 = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1$
  - $E(Y/\text{Variedad } 2) = \beta_0 + \beta_1 X_1 + 0 + \beta_3 + 0 + \beta_5 X_1 = (\beta_0 + \beta_2) + (\beta_3 + \beta_5) X_1$

- Por tanto la diferencia entre el Rto medio con la Variedad 2 respecto al obtenido cultivando la Variedad será $= \beta_2 + \beta_4 X_1$, que depende de la dosis de abonado (si $\beta_4 \neq 0$), e igual sucede con la diferencia entre la Variedad 3 y la Variedad 1 que resulta ser $\beta_3 + \beta_5 X_1$.

- Desde otro punto de vista, y como se aprecia en la figura, el modelo contempla la posibilidad, si β_4 o β_5 son diferentes de cero, de que el efecto de la dosis de abonado sea diferente según la variedad considerada (o sea de que las pendientes de las rectas sean diferentes).


E(Rto)
Variedad 3
$\beta_1+\beta_5$
Variedad 2
$\beta_0+\beta_3$
$\beta_1+\beta_4$
$\beta_0+\beta_2$
Variedad 1
β_1
β_0
DOSIS

Ejemplo de la utilización del Modelo de Regresión Clásico: Comparación de dos fármacos contra la hipertensión

Se recogen a continuación los datos de un estudio para comparar un nuevo fármaco B contra la hipertensión con un fármaco tradicional A. De una muestra de 60 pacientes hipertensos, 30 seleccionados al azar se trataron con A y los otros 30 con B. Los datos recogen, además del sexo del paciente, la tensión (mm Hg) antes del tratamiento y tras 3 semanas del inicio del mismo (las tensiones son, en ambos casos, las medias de 3 determinaciones diarias en 3 días consecutivos).

	Farmaco A			Farmaco B	
SEXO	tens_antes	tens_despues	SEXO	tens_antes	tens_despues
V	200	160	V	181	132
M	193	151	V	162	141
M	169	148	M	194	131
M	183	147	V	176	139
M	181	142	M	156	132
V	179	151	V	175	138
M	190	146	V	189	139
M	185	147	V	166	143
M	195	151	M	163	128
M	206	161	M	193	138
M	188	146	V	195	148
V	203	170	V	174	140
V	197	168	V	191	149
M	201	157	M	181	136
M	207	159	M	198	128
M	206	157	V	193	150
V	191	160	M	185	131
M	183	138	V	184	141
M	212	165	V	192	147
M	187	147	M	206	138
M	201	154	V	214	157
M	182	146	V	174	146
M	213	151	M	174	124
V	198	164	M	178	136
V	188	158	M	191	131
V	214	177	M	172	131
M	186	143	M	187	126
M	165	135	M	217	141
M	170	134	V	192	148
M	175	134	M	203	139

¿Cómo se opera mediante Modelos de Regresión para extraer la información que hay en estos datos sobre las diferencias entre los dos fármacos?

Test t de Student de comparación de 2 medias

- En principio, el estudio de la diferencia entre los dos fármacos podría abordarse de forma sencilla mediante el clásico test t de Student de comparación de 2 medias.
- En efecto para la variable de interés: diferencia entre la tensión antes y después del tratamiento, se dispone de 30 observaciones en la población de pacientes tratados con A y de otras 30 en la de los pacientes tratados con B.
- El test t de Student de comparación de 2 medias da los siguientes resultados:

Comparación de Medias para bajada

Prueba t para comparar medias
Hipótesis nula: media1 = media2
Hipótesis Alt.: media1 <> media2
suponiendo varianzas iguales: t = -1.71483 valor-P = 0.0917148

Intervalos de confianza del 95.0% intervalo de confianza para la diferencia de medias
suponiendo varianzas iguales: -5.06667 +/- 5.91431 [-10.981; 0.847643]

- Se estima, por tanto, que la bajada de tensión es 5.067 mmHg mayor utilizando B que utilizando A, pero esta diferencia no llega a ser significativa estadísticamente (para α =0.05) por ser p-value = 0.092

Comparación de dos medias mediante un modelo de regresión

- El test t de comparación de medias no es más que un caso particular, especialmente sencillo, de un Modelo de Regresión Clásico
- En efecto, si se plantea un modelo con la bajada de tensión como variable dependiente y la variable explicativa X_B, que valga 0 cuando se administró el fármaco A y 1 cuando se administró el fármaco B:
- $$\mathbf{E(bajada) = \beta_0 + \beta_1 X_B} \quad (1)$$
- La interpretación de los parámetros de este modelo será:
  - $\boldsymbol{\beta_0}$ = E(bajada/X_B=0) = E(bajada) cuando se da el fármaco A
  - $\boldsymbol{\beta_0 + \beta_1}$ = E(bajada/X_B=1) = E(bajada) cuando se da el fármaco B
  - $\boldsymbol{\beta_1}$ = diferencia en la bajada media de tensión si se utiliza B en vez de A
- Por lo tanto, cuantificar la diferencia de medias entre los fármacos B y A y analizar su significación estadística, es equivalente a estimar el parámetro $\boldsymbol{\beta_1}$ en el Modelo de Regresión (1) y estudiar si dicho parámetro es significativo

Comparación de dos medias mediante un modelo de regresión (continuación)

- En efecto, si se ajusta el Modelo de Regresión (1) a los datos anteriores se obtiene:

Parámetro	*Estimación*	*Error Estándar*	*Estadístico T*	*Valor-P*
CONSTANTE	40.1	2.08923	19.1937	0.0000
XB	5.06667	2.95461	1.71483	0.0917

Error estándar del est. = 11.4432

- Puede constatarse que, como era lógico, la estimación de la diferencia de medias (5.067) y la significación de dicha diferencia (0.092) resultan idénticos a los resultados obtenidos mediante el test t de Student.
- ¡Pero el enfoque mediante Modelos de Regresión no es sólo una alternativa equivalente al clásico test t de Student, sino que, como se expone a continuación, permite análisis mucho más potentes e informativos!

Limitaciones del enfoque clásico

- En principio cuando se comparan dos muestras, éstas deberían ser lo más parecidas posibles respecto a todos los factores, diferentes del estudiado, que pueden influir sobre la respuesta.

- En el ejemplo, es posible que la bajada de tensión sea mayor cuanto más alta sea la tensión inicial, o que sea mayor en un sexo que en otro. Ello plantea tres problemas:

  - Si las dos muestras no están perfectamente equilibradas respecto a sexos y a valores iniciales de las tensiones (lo que puede ser difícil de lograr), las diferencias al respecto pueden sesgar la estimación de la diferencia entre fármacos

  - Aunque las dos muestras estuvieran perfectamente equilibradas respecto a dichos factores, el efecto de éstos incrementará la variabilidad dentro de cada muestra, reduciendo la potencia del estudio para detectar como significativa la diferencia entre fármacos.

  - Con este enfoque no resulta posible analizar si la diferencia entre fármacos depende de alguno de estos otros factores (posible existencia de interacciones)

Ventajas del enfoque mediante Modelos de Regresión

- Frente a estos problemas, resulta muy sencillo incluir en el MRC variables adicionales asociadas a la tensión inicial y al sexo. Ello permitirá:

- Obtener una **estimación insesgada** de la diferencia entre fármacos, aunque las muestras no estén perfectamente equilibradas respecto a sexos y tensión inicial.

- Obtener una **estimación más precisa** de dicha diferencia, puesto que la variabilidad debida al efecto de ambos factores estará recogida en los parámetros correspondientes y no incrementará la variabilidad residual.

- Obtener adicionalmente una **estimación del efecto de los otros factores** (sexo y tensión inicial) sobre la bajada de tensión.

- Introduciendo, como veremos, nuevas variables XB*sexo y XB*tension_inicial, **estudiar si hay interacciones**, o sea si la diferencia entre fármacos depende significativamente del sexo o de la tensión inicial

Inclusión en el modelo del efecto del sexo y la tensión inicial

- Con el fin de eliminar el posible sesgo en la estimación de la diferencia entre fármacos y obtener una estimación más precisa de dicha diferencia, es posible (y muy sencillo) incluir en el modelo los posibles efectos del sexo y de la tensión inicial
- $$\mathbf{E(bajada) = \beta_0 + \beta_1 X_B + \beta_2 X_M + \beta_3 X_{TI}}$$
- donde: X_M: variable que vale 1 en mujeres y 0 en varones
- X_{TI}: tensión inicial (mm Hg por encima de 180)
- La interpretación de los parámetros en el nuevo modelo es ahora:
- β_0 : bajada de tensión media con el fármaco A (X_B=0) cuando sexo es varón (X_M =0) y la tensión inicial es 180 (X_{TI} =0)
- β_2 : diferencia en la bajada de tensión media entre mujeres y varones (se asume que esa diferencia es la misma con los dos fármacos y no depende de X_{TI})
- β_3 :aumento en la bajada de tensión media por cada mmHg de aumento en la tensión inicial (se asume que no depende del fármaco ni del sexo)
- β_1 : diferencia en la bajada de tensión media entre fármaco B y fármaco A cuando se administran a pacientes del mismo sexo y tensión inicial (se asume que esa diferencia no depende del sexo ni de X_{TI})

Conclusiones obtenidas con el nuevo modelo

- El ajuste del nuevo modelo proporciona los siguientes resultados:

Parámetro	*Estimación*	*Error Estándar*	*Estadístico T*	*Valor-P*
CONSTANTE	24.0	1.32235	18.1495	0.0000
X_B	13.0536	1.25795	10.3768	0.0000
X_M	12.6791	1.24286	10.2016	0.0000
X_{TI}	0.584704	0.0401884	14.5491	0.0000

Error estándar del est. = 4.54355

CONCLUSIONES:

Los fármacos son más efectivos en mujeres que en hombres (β_2=12.7 muy significativa estadísticamente)

La bajada de tensión es tanto mayor cuanto mayor es la tensión inicial (β_3=.58 muy significativa estadísticamente)

La estimación de la diferencia entre fármacos es **13.05 mmHg, resultando muy significativa estadísticamente**. La diferencia (5.07) obtenida en el estudio sencillo subestimaba el valor real, al no tener en cuenta que en el grupo tratado con B había una proporción algo menor de mujeres y la tensión inicial era en promedio algo más baja.

El modelo obtenido es más preciso (s_{resid}=4.54 frente a 11.4 en el modelo simplificado) al haber eliminado de la misma los efectos de la variabilidad entre sexos y entre tensiones iniciales dentro de ambas

Análisis de las interacciones

- Es posible profundizar todavía algo más la investigación, estudiando la posible presencia de interacciones entre los efectos de los 3 factores estudiados (fármaco, sexo y tensión inicial)
- Basta para ello, **y es muy sencillo**, incluir en el modelo tres nuevas variables definidas como productos entre las 3 parejas posibles de variables

$$E(bajada) = \beta_0 + \beta_1 X_B + \beta_2 X_M + \beta_3 X_{TI} + \beta_4 X_B X_M + \beta_5 X_B X_{TI} + \beta_6 X_M X_{TI}$$

- ¿Qué interpretación tienen en este nuevo modelo los parámetros asociados a la diferencia entre fármacos?
- β_1 : diferencia en la bajada de tensión media entre fármaco B y fármaco A cuando se administran a varones con tensión inicial = 180
- β_4 : aumento de la diferencia entre los dos fármacos cuando se administran a mujeres en vez de a varones
- β_5 : aumento de la diferencia entre los dos fármacos cuando aumenta 1 mmHg la tensión inicial

Análisis de las interacciones (continuación)

- El ajuste del nuevo modelo proporciona, inmediatamente, los siguientes resultados:

Parámetro	*Estimación*	*Error Estándar*	*Estadístico T*	*Valor-P*
CONSTANTE	28.2423	1.8612	15.1742	0.0000
XB	8.90737	2.07258	4.29773	0.0001
XM	10.4955	1.93645	5.41998	0.0000
XTI	0.300476	0.0816949	3.67802	0.0006
XB*XM	0.843226	2.26919	0.371598	0.7117
XB*XTI	0.396342	0.0709278	5.58796	0.0000
XM*XTI	0.0911643	0.0751214	1.21356	0.2303

- **CONCLUSIONES**
- Las interacciones entre el sexo y los otros dos factores no son significativas
- Existe una **interacción muy significativa** entre el efecto de los fármacos y la tensión inicial
- Según los parámetros estimados, la diferencia entre la bajada media de la tensión al aplicar B en vez de A es: $8.907 + 0.396X_{TI}$
- Así, la diferencia estimada es 24.7 (8.9 + 0.396x40) si la tensión inicial es 220 mmHg, y es prácticamente nula (8.9-0.396x20 = 0.98) si la tensión inicial es sólo de 160 mmHg.

¿Qué información han extraído de los datos los Modelos de Regresión?

Fármaco A			Fármaco B		
SEXO	tens_antes	tens_despues	SEXO	tens_antes	tens_despues
V	200	160	V	151	132
M	193	151	V	162	141
M	189	140	M	194	131
M	193	147	V	175	139
M	181	142	M	158	132
V	179	151	V	176	138
M	190	146	V	169	139
M	185	147	V	166	143
M	195	151	M	168	129
M	206	161	M	190	138
M	198	146	V	190	149
V	203	170	V	174	140
V	197	165	V	191	140
M	201	157	M	191	136
M	207	159	M	190	128
M	205	157	V	193	150
V	191	160	M	180	131
M	183	139	M	184	141
M	212	166	V	182	147
M	187	147	M	206	138
M	201	154	V	214	157
M	182	149	V	174	146
M	219	151	M	174	124
V	198	154	M	178	135
V	189	158	M	191	131
V	214	177	M	172	131
M	188	143	M	167	126
M	195	125	M	217	141
M	170	134	M	192	146
M	175	134	M	203	139

- **Conclusiones mediante el test t**: la diferencia entre B y A es en promedio 5.07 mmHg, no estando claro si es estadísticamente significativa (p-val: 0.092)
- **Conclusiones utilizando Modelos de Regresión**:
  - La diferencia entre B y A es en promedio 13.05 mmHg, resultando muy significativa estadísticamente (p-val: 0.0000)
  - La diferencia media entre B y A es tanto mayor cuanto más alta es la tensión inicial, siendo del orden de 25 si la tensión inicial es 220 mmHg, y prácticamente nula si la tensión inicial es sólo de 160 mmHg
  - Ambos fármacos son más efectivos en mujeres que en hombres, siendo la diferencia de efectividad del orden de 10 mmHg y no pareciendo depender del fármaco ni de la tensión inicial

Un test general para el análisis de Modelos de Regresión

- Frecuentemente en Modelos de Regresión se desea investigar la aceptabilidad de hipótesis que implican restricciones sobre varios parámetros del modelo
- Por ejemplo, en el modelo para estudiar el efecto de 3 variedades y la dosis de abonado sobre el rendimiento de un cultivo:
- $$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 X_1 Z_2 + \beta_5 X_1 Z_3$$
- Hipótesis: No hay interacciones ←→ $\beta_4=0$ y $\beta_5=0$ (2 restricciones)
- Hipótesis: Variedad 1 = Variedad 2 ←→ $\beta_2=\beta_3$ y $\beta_4=\beta_5$ (2 restricciones)
- Un procedimiento general para ver si es admisible una hipótesis que implica determinadas restricciones sobre los parámetros del modelo consiste en estudiar si el incremento de la $SC_{residual}$ que se produce al forzar que el ajuste satisfaga estas restricciones es, o no, estadística-mente significativo
- Dicho Δ $SC_{residual}$ dividido por el núero de restricciones impuestas en los parámetros, se compara con el $CM_{residual}$ del modelo completo (sin restricciones) mediante un test F, tal como se ve en la siguiente diapositiva

Operativa del test general

	Suma de Cuadrados Residual	Grados de Libertad Residuales
Modelo Restringido	SCR_r	N - 1 - I + r
Modelo Completo	SCR_c	N - 1 - I
Incremento	Δ SCR = SCR_r - SCR_c (siempre es $\exists$ 0)	r

- Se demuestra que si es cierta la H_0 que implican las r restricciones:

$$\frac{\Delta SCR / r}{SCR_c / (N-1-I)}$$ se distribuye como una $F_{r,(N-1-I)}$

- mientras que si H_0 es falsa, el cociente anterior resulta mayor en promedio que una $F_{r,N-1-I}$. La H_0 se rechaza por tanto (con riesgo de 1ª especie α) si el cociente resulta mayor que $F_{r,N-1-I}(\alpha)$

- Ejercicio: en el ejemplo de las 3 variedades y la dosis ¿cómo se operaria en Statgraphics para obtener la SCresidual del modelo cuando se imponen las siguientes restricciones?
  - 1) $\beta_4=0$ y $\beta_5=0$ 2) $\beta_2=\beta_3$ y $\beta_4=\beta_5$

Ejemplo: efecto del día de la semana en el consumo de energía

- Con el fin de mejorar las predicciones del modelo se realizó un nuevo ajuste incluyendo entre las variables explicativas la temperatura de la víspera y el día de la semana (variable cualitativa con 5 variantes modelizada mediante 4 "dummys" asociadas a martes,…, viernes)
- $\mathbf{E(Consumo) = ß_0 + ß_1 Temp + ß_2 Temp^2 + ß_3 T\_ant + ß_4 Ma + ß_5 Mi + ß_6 Ju + ß_7 Vi \quad (1)}$
- donde, p.e., "Ma" vale 1 los martes y 0 el resto de los días (Nota: en Statgraphics se puede introducir esta variable poniendo "DIA=2")
- ¿Qué son los diferentes β_i en el modelo anterior?
- Los resultados del ajuste del Modelo 1 (datos en *gas_2.sgd)* son los siguientes

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	478.805	9.74108	49.1531	0.0000
Temper	-22.8388	2.42179	-9.43052	0.0000
Temper^2	0.353422	0.0926083	3.81632	0.0004
Temp_ant	-2.37723	1.4228	-1.67081	0.1011
DIA=2	-2.05894	9.83595	-0.209328	0.8351
DIA=3	-10.994	8.31683	-1.3219	0.1923
DIA=4	-0.297069	9.10169	-0.032639	0.9741
DIA=5	-22.1611	8.06962	-2.74624	0.0084

¿Es significativo el efecto del día de la semana?

- El cuadro resumen del Anova del Modelo (1) es:

```
                          Analysis of Variance
--------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
--------------------------------------------------------------------------------
Model                     566823.0      7      80974.7     193.94       0.0000
Residual                   20458.9     49      417.529
--------------------------------------------------------------------------------
Total (Corr.)             587282.0     56
```

- Si se ajusta un nuevo modelo (Modelo 2) excluyendo las 4 variables asociadas al día de la semana, el cuadro del Anova es:

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	562393.	3	187464.	399.14	0.0000
Residuo	24892.6	53	469.672		
Total (Corr.)	587285.	56			

- ¿Es significativo el ΔSC_{resid} al forzar $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$?

$$\frac{(24892.6 - 20489.9)/4}{417.53} = 2.65 \quad > \quad F_{4,49}^{0.05} = 2.56$$

No es admisible la Hipótesis $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

¿Hay diferencias significativas de lunes a jueves?

- Los resultados del Modelo 1 sugieren que sólo es significativa la diferencia entre el viernes y el resto de los días
- Para estudiar si es admisible la H_o: $\beta_4 = \beta_5 = \beta_6 = 0$ se ajusta el modelo:
  - $$E(Consumo) = ß_0 + ß_1 Temp + ß_2 Temp^2 + ß_3 T\_ant + ß_7 Vi \quad (3)$$
- ¿Qué interpretación tienen los β_i de este Modelo 3?

		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	475.305	7.94481	59.8259	0.0000
TEMPER	-22.3307	2.20002	-10.1502	0.0000
TEMPER^2	0.355977	0.0864741	4.11658	0.0001
TEMPANT	-2.97119	1.30453	-2.27759	0.0269
DIA=5	-18.6485	6.40158	-2.91311	0.0053

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	565885.	4	141471.	343.76	0.0000
Residuo	21400.2	52	411.542		
Total (Corr.)	587285.	56			

- ¿Es admisible la hipótesis de que sólo el viernes difiere de los restantes días?. Sabiendo que el consumo de cada día se mide desde que entra el primer turno a las 6:30 hasta que entra el del día siguiente ¿a qué puede deberse el resultado obtenido?
- Estudiar si la diferencia entre el viernes y los restantes días es tanto mayor cuanto más frío hace

Relación entre los Modelos de Regresión y el ANOVA

- El Anova es, en el fondo, un caso particular de los modelos de regresión
- La SC asociada en un Anova a cualquier factor, o interacción, no es más que el ΔSC_{resid} cuando se eliminan del modelo las variables dummy asociadas a ese efecto
- **Ejercicio**
- En el archivo *regresion_anova.sgd* se recogen los resultados de un estudio sobre la influencia de la variedad (2 variantes) y el tipo de suelo (3 variantes) sobre el rendimiento de un cultivo (plan 2x3 con 2 réplicas)
  - Analizar los resultados mediante un Anova (incluir la interacción)
  - Ajustar los datos a un Modelo de Regresión, incluyendo una dummy para la variedad, dos dummys para el tipo de suelo y otras dos dummys para la interacción
  - Comprobar que la SC_{resid} del ajuste es exactamente igual a la del Anova y tiene los mismos grados de libertad
  - Repetir el ajuste sin incluir las dos variables asociadas a la interacción. Comprobar que el ΔSC_{resid} coincide exactamente con la SC asociada a la interacción en el Anova

Anova en modelos desequilibrados: medias "least squares"

- Al realizar un Anova, Statgraphics proporciona la tabla de medias "least squares" para las variantes de los diferentes factores (y para sus combinaciones si se incluyen interacciones)
- Cuando el Anova es desequilibrado (distinto número de datos en cada casilla) estas medias "least squares" no coinciden siempre con los valores medios calculados directamente de los datos
- Para comprender lo que son estas medias "least squares" se considera el siguiente conjunto de datos para un Anova 2x2. (los datos vienen en negro, las medias de las casillas en azul y las medias marginales en rojo)

	B1	B2	
A1	22 18 $m_{11} = 20$	39 41 $m_{12} = 40$	$m_{A1} = 30$
A2	19 21 $m_{21} = 20$	38 40 40 42 41 39 39 41 $m_{22} = 40$	$m_{A2} = 36$
	$m_{B1} = 20$	$m_{B2} = 40$	

medias "least squares" (continuación)

- En la tabla se aprecia un efecto claro del factor B, siendo m_{B2}=40 más elevado que que m_{B1}=20. Esta diferencia de 20 entre los valores medios de B se mantiene tanto si A=A1 como si A=A2 (la interacción es nula)
- Aparentemente también hay un efecto del factor A, puesto que m_{A2}=36 es más elevado que m_{A1}=30.
- Sin embargo es evidente que A no tiene ningún efecto, porque $m_{11} = m_{21} = 20$ y $m_{12} = m_{22} = 40$, o sea que la media de A1 es igual a la media de A2 tanto si B=1 como si B=2
- ¿Cómo es que, sin embargo, se ha obtenido m_{A2} más elevado que m_{A1}?
- ¿Cuáles hubieran sido m_{A1} y m_{A2} si hubiera habido el mismo número de datos en todas las casillas?
- Analizar mediante Statgraphics los datos de la tabla (datos en *anova_desq.sgd*) y comprobar
  - Que las SC asociadas al efecto de A y a la interacción son nulas (como debe ser)
  - Que las medias "least squares" de A1 y de A2 son ambas iguales a 30, que sería la media esperada en ambos casos para el promedio de las dos alternativas para B
  - ¿Cómo habrá obtenido estos resultados Statgraphics?