CAPíTULO 7

Regresores deterministas

7.1. Variables ficticias

Los datos estadísticos que se utilizan en un modelo de regresión pueden corresponder a variables cuantitativas y cualitativas. Mientras que las modalidades de una variable cuantitativa son medibles, las de una variable cualitativa no lo son. Por ejemplo, en una muestra de personas, las modalidades de la variable cualitativa sexo son: *femenino* y *masculino*, que no son medibles. Ahora bien, estas modalidades **inducen una clasificación** de las personas de la muestra en dos grupos, mujeres y hombres, y podemos definir una variable artificial o ficticia M_i que toma el valor 1 para indicar que la persona i es mujer y el valor 0 para indicar que es hombre, y una variable ficticia H_i que toma el valor 1 para indicar que la persona i es hombre y el valor 0 para indicar que es mujer

$$M_i = \begin{cases} 1 & \text{si } i \in \text{mujer} \\ 0 & \text{si } i \in \text{hombre} \end{cases} \qquad H_i = \begin{cases} 1 & \text{si } i \in \text{hombre} \\ 0 & \text{si } i \in \text{mujer} \end{cases}$$

Análogamente, la variable cualitativa estado civil induce una clasificación de las personas en la muestra en tres grupos: solteros, casados y viudos. Así, podríamos definir una variable ficticia E_i que toma los valores 0, 1 y 2 si la persona i es soltera, casada o viuda, respectivamente. Alternativamente, podemos definir una variable binaria para cada modalidad del estado civil. Así, la variable S_i toma el valor 1 si la persona i pertenece al grupo de los solteros, y 0 en caso contrario; la variable C_i toma el valor 1 si la persona i pertenece al grupo de los casados, y 0 en caso contrario; y la variable V_i toma el valor 1 si la persona i pertenece al grupo de los viudos, y 0 en caso contrario:

$$S_i = \begin{cases} 1 & \text{si } i \in \text{solteros} \\ 0 & \text{si } i \notin \text{solteros} \end{cases} \quad C_i = \begin{cases} 1 & \text{si } i \in \text{casados} \\ 0 & \text{si } i \notin \text{casados} \end{cases} \quad V_i = \begin{cases} 1 & \text{si } i \in \text{viudos} \\ 0 & \text{si } i \notin \text{viudos} \end{cases}$$

En ocasiones puede ser conveniente definir varias variables binarias a partir de una variable cuantitativa. Por ejemplo, las observaciones de la variable cuantitativa renta disponible R para una muestra de familias ($i = 1, \ldots, n$) pueden clasificarse en tres grupos: renta baja RB, renta media RM y renta alta RA. Fijados tres umbrales de renta a, b y c, clasificamos una familia en el grupo RB si $R_i < a$; en el grupo RM, si $a < X_i < b$; y en el grupo RA, si $X_i > b$. De aquí, definimos las variables ficticias

$$RB_i = \begin{cases} 1 & \text{si } i \in RB \\ 0 & \text{si } i \notin RB \end{cases} \quad RM_i = \begin{cases} 1 & \text{si } i \in RM \\ 0 & \text{si } i \notin RM \end{cases} \quad RA_i = \begin{cases} 1 & \text{si } i \in RA \\ 0 & \text{si } i \notin RA \end{cases}$$

Definición 48. *Una variable discreta es binaria, dicotómica o dummy cuando toma sólo dos valores (0 ó 1) y es policotómica cuando toma más de dos valores.*

EJERCICIO 1. *Considere la lista de calificaciones en el examen final de econometría. Especifique las posibles modalidades de las siguientes variables cualitativas o, equivalentemente, los grupos inducidos por las mismas: (1) el alumno cursó la asignatura optativa de inferencia estadística, (2) el alumno cursa por primera vez la asignatura, (3) el alumno asiste a las clases de teoría y/o práctica. ¿Cabe esperar alguna diferencia en la calificación media de cada grupo atribuible a estas variables cualitativas?*

Observación 35. *Las variables binarias o dummy d_{ji} asociadas a todas las modalidades $j = 1, \ldots, m$ de una variable cualitativa cumplen la restricción:*

$$d_{1i} + d_{2i} + \cdots + d_{mi} = 1$$

indicando que cada observación i pertenece a uno y sólo uno de los m grupos inducidos por la variable cualitativa.

Los métodos econométricos presentados hasta ahora son válidos cuando el modelo de regresión expresa una variable dependiente cuantitativa en función de variables explicativas cuantitativas y cualitativas. Sin embargo, los modelos de regresión con variable dependiente cualitativa requieren el desarrollo de métodos econométricos específicos que no son objeto de estudio en este curso. A continuación se consideran diversas aplicaciones de las variables explicativas ficticias en el análisis de regresión.

7.2. Modelo de regresión con una variable cualitativa

7.2.1. Variable cualitativa con dos modalidades. Un problema estadístico clásico es la comparación de las medias de dos distribuciones normales. Supongamos que las n observaciones $Y_1, Y_2, \ldots, Y_n$ provienen de dos distribuciones normales con medias μ_1 y μ_2 y varianza común σ^2. En concreto, $Y_i \sim iidN(\mu_1, \sigma^2)$ para $i = 1, \ldots, n_1$ e $Y_i \sim iidN(\mu_2, \sigma^2)$ para $i = n_1 + 1, \ldots, n$. Vemos que podemos formar dos grupos de observaciones en la muestra: el primero contiene las primeras n_1 observaciones; y el segundo, las $n_2 = n - n_1$ restantes observaciones. Queremos contrastar la hipótesis $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ al nivel de significación α.

EJEMPLO 12. *Supongamos que observamos el salario de n personas con idéntica categoría laboral: $Y_1, Y_2, \ldots, Y_n$. Si las n_1 primeras observaciones corresponden a mujeres y las n_2 últimas observaciones a hombres, la distribución salarial para las mujeres es $Y_i \sim iidN(\mu_1, \sigma^2)$ y para los hombres $Y_i \sim iidN(\mu_2, \sigma^2)$. Deseamos contrastar si el salario medio para las mujeres μ_1 es igual que el salario medio para los hombres μ_2.* ◇

Podemos formular el contraste de igualdad de medias en el marco del modelo lineal general. Así, bajo H_0 tenemos el modelo de regresión con término constante

$$Y_i = \mu + u_i, \quad i = 1, \ldots, n$$

en donde $u_i \sim iidN(0, \sigma^2)$. El estimador de mínimos cuadrados de la ordenada es la media muestral de la variable dependiente

$$\hat{\mu} = \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

y su varianza

$$V(\hat{\mu}) = \frac{\sigma^2}{n}$$

Bajo H_1 tendríamos una ecuación de regresión para cada uno de los dos grupos de observaciones

$$\begin{aligned} Y_i =& \mu_1 + u_i, \quad i = 1, \ldots, n_1 \\ Y_i =& \mu_2 + u_i, \quad i = n_1 + 1, \ldots, n \end{aligned} \tag{7.1}$$

siendo los estimadores mínimo-cuadráticos de μ_1 y μ_2

$$\hat{\mu}_1 = \bar{Y}_1 = \frac{\sum_{i=1}^{n_1} Y_i}{n_1} \quad \text{y} \quad \hat{\mu}_2 = \bar{Y}_2 = \frac{\sum_{i=n_1+1}^{n} Y_i}{n - n_1}$$

y sus varianzas

$$V(\hat{\mu}_1) = \frac{\sigma^2}{n_1} \quad \text{y} \quad V(\hat{\mu}_1) = \frac{\sigma^2}{n - n_1}$$

Las dos ecuaciones de regresión en (7.1) pueden combinarse en una ecuación de regresión múltiple mediante el empleo de variables dummy

$$Y_i = \mu_1 d_{1i} + \mu_2 d_{2i} + u_i \tag{7.2}$$

en donde d_{1i} y d_{2i} son dos variables *dummy* definidas del siguiente modo

$$d_{1i} = \begin{cases} 1 & \text{si } i \in \{1, \ldots, n_1\} \\ 0 & \text{si } i \in \{n_1 + 1, \ldots, n\} \end{cases} \qquad d_{2i} = \begin{cases} 0 & \text{si } i \in \{1, \ldots, n_1\} \\ 1 & \text{si } i \in \{n_1 + 1, \ldots, n\} \end{cases}$$

Vemos que cuando la observación i proviene de la primera distribución $N(\mu_1, \sigma^2)$, $d_{1i} = 1$ y $d_{2i} = 0$, la ecuación (7.2) se reduce a $Y_i = \mu_1 + u_i$; mientras que cuando la observación i proviene de la segunda distribución $N(\mu_2, \sigma^2)$, $d_{1i} = 0$ y $d_{2i} = 1$, la ecuación (7.2) se reduce a $Y_i = \mu_2 + u_i$.

El estimador de mínimos cuadrados en (7.2) es

$$\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} d_{1i}^2 & \sum_{i=1}^{n} d_{1i} d_{2i} \\ \sum_{i=1}^{n} d_{1i} d_{2i} & \sum_{i=1}^{n} d_{2i}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} d_{1i} Y_i \\ \sum_{i=1}^{n} d_{2i} Y_i \end{pmatrix}$$

Ahora bien, de los n valores de la variable d_{1i}, n_1 son iguales a 1 y $n - n_1$ son iguales a 0; por tanto, la suma de los cuadrados de d_{1i} es n_1. Del mismo modo, la suma de los cuadrados de d_{2i} es igual a $n - n_1$. Además, cuando $d_{1i} = 1$ se tiene que $d_{2i} = 0$, y viceversa; por tanto, la suma de los productos cruzados es cero. Finalmente, la suma de los productos cruzados de d_{1i} e Y_i es el total de Y para el primer grupo, $\sum_{i=1}^{n_1} Y_i$, y la correspondiente a d_{2i} e Y_i es el total de Y para el segundo grupo, $\sum_{i=n_1+1}^{n} Y_i$. De aquí,

$$\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n_1} Y_i \\ \sum_{i=n_1+1}^{n} Y_i \end{pmatrix} = \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 \end{pmatrix}$$

Observación 36. *Como las variables dummy d_{1i} y d_{2i} son ortogonales, las estimaciones en la regresión multiple (7.2) coinciden con las obtenidas en las regresiones simples (7.1).*

Para contrastar la hipótesis $H_0 : \mu_1 = \mu_2$ frente a la alternativa $H_1 : \mu_1 \neq \mu_2$, podemos utilizar el estadístico t

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{V}(\hat{\mu}_1 - \hat{\mu}_2)}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\dfrac{\hat{\sigma}^2}{n_1} + \dfrac{\hat{\sigma}^2}{n - n_1}}} \sim t_{n-2}$$

en donde

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2} = \frac{\sum_{i=1}^{n_1}(Y_i - \bar{Y}_1)^2 + \sum_{i=n_1+1}^{n}(Y_i - \bar{Y}_2)^2}{n-2}$$

La hipótesis $H_0 : \mu_1 = \mu_2$ se rechaza al nivel de significación α si $|t| > c$, en donde c es el valor crítico tal que $Prob(-c < t_{n-k} < c) = 1 - \alpha$.

Ejercicio 2. *Demuestre que el estadístico F para contrastar la restricción lineal $H_0 : \mu_1 - \mu_2 = 0$ en (7.2) es equivalente al estadístico t.*

La relación $d_{1i} + d_{2i} = 1$ nos permite reescribir la ecuación (7.2) en términos de una única variable explicativa, d_{1i} o d_{2i}. Así, reemplazando d_{2i} por $1 - d_{1i}$ tenemos

$$Y_i = \mu_1 d_{1i} + \mu_2(1 - d_{1i}) + u_i = \mu_2 + (\mu_1 - \mu_2)d_{1i} + u_i$$

Esta ecuación se lee del siguiente modo: regresión simple de Y_i sobre d_{1i}, cuya representación general es

$$Y_i = \alpha_1 + \alpha_2 d_{1i} + u_i$$

en donde la ordenada $\alpha_1 = \mu_2$ y la pendiente $\alpha_2 = \mu_1 - \mu_2$. Por tanto, en la regresión simple de Y_i sobre d_{1i}, la ordenada estimada es la media de la variable dependiente para el grupo 2, $\bar{Y}_2$, y la pendiente de d_{1i} es la diferencia entre las medias de la variable dependiente para el grupo 1 y el grupo 2, $\bar{Y}_1 - \bar{Y}_2$. Note que la variable ficticia omitida, d_{2i}, determina el **grupo base** respecto al que se hacen las comparaciones.

De acuerdo con lo anterior, en la regresión simple de Y_i sobre d_{2i}

$$Y_i = \delta_1 + \delta_2 d_{2i} + u_i$$

el grupo base es el 1; la ordenada estimada es la media de la variable dependiente para el grupo base, $\hat{\delta}_1 = \bar{Y}_1$; y la pendiente estimada de d_{2i} es la diferencia entre las medias de la variable dependiente para el grupo 2 y el grupo 1, $\hat{\delta}_2 = \bar{Y}_2 - \bar{Y}_1$. Es claro que esta ecuación se obtiene sustituyendo en (7.2) la variable ficticia d_{1i} por $1 - d_{2i}$

$$Y_i = \mu_1(1 - d_{2i}) + \mu_2 d_{2i} + u_i = \mu_1 + (\mu_2 - \mu_1)d_{2i} + u_i$$

El contraste de igualdad de medias, $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$, en la ecuación $Y_i = \delta_1 + \delta_2 d_{2i} + u_i$ es simplemente un contraste de significiación individual $H_0 : \delta_2 = 0$ versus $H_1 : \delta_2 \neq 0$ que puede basarse en la ratio t

$$t = \frac{\hat{\delta}_2}{\sqrt{\hat{V}(\hat{\delta}_2)}} = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\hat{V}(\hat{\mu}_2 - \hat{\mu}_1)}} \sim t_{n-2}$$

Proposición 61. *Si en la ecuación de regresión (7.2) se incluye un término constante, entonces se introduce multicolinealidad exacta. Este problema se denomina* **la trampa de las variables ficticias**.

Demostración. En la ecuación (7.2) con término constante

$$Y_i = \beta_0 + \beta_1 d_{1i} + \beta_2 d_{2i} + u_i, \quad i = 1, \ldots, n$$

la variable asociada al termino constante, $x_{1i} = 1$, es una combinación lineal de d_{1i} y d_{2i}, $d_{1i} + d_{2i} = 1$. De aquí, el estimador de mínimos cuadrados es

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} n & n_1 & n_2 \\ n_1 & n_1 & 0 \\ n_2 & 0 & n_2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n d_{1i} Y_i \\ \sum_{i=1}^n d_{2i} Y_i \end{pmatrix}$$

Vemos que la matriz $\mathbf{X'X}$ es singular: la primera columna es igual a la segunda más la tercera. □

Observación 37. *Como veremos en otro tema, el problema de multicolinealidad exacta puede evitarse imponiendo una restricción sobre los parámetros. Por ejemplo, si fijamos $\beta_0 = \bar{Y}$, entonces tenemos la regresión*

$$Y_i - \bar{Y} = \beta_1 d_{1i} + \beta_2 d_{2i} + u_i$$

siendo los estimadores mínimo-cuadráticos $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}$ y $\hat{\beta}_2 = \bar{Y}_2 - \bar{Y}$ las diferencias entre la media de cada grupo y la media global.

En resumen, la comparación de las medias de dos distribuciones normales puede basarse en las siguientes regresiones:

1. $Y_i = \alpha_1 d_{1i} + \alpha_2 d_{2i} + u_i$
2. $Y_i = \beta_1 + \beta_2 d_{2i} + u_i$

en donde $\hat{\alpha}_1 = \bar{Y}_1$, $\hat{\alpha}_2 = \bar{Y}_2$, $\hat{\beta}_1 = \bar{Y}_1$, $\hat{\beta}_2 = \bar{Y}_2 - \bar{Y}_1$.

7.2.2. Variable cualitativa con múltiples modalidades. El análisis anterior se extiende fácilmente a la comparación de las medias de tres o más distribuciones normales. Como ilustración suponemos que las n observaciones $Y_1, Y_2, \ldots, Y_n$ provienen de tres distribuciones normales con medias μ_1, μ_2 y μ_3 y varianza común σ^2. En concreto, $Y_i \sim iidN(\mu_1, \sigma^2)$ para $i = 1, \ldots, n_1$, $Y_i \sim iidN(\mu_2, \sigma^2)$ para $i = n_1 + 1, \ldots, n_1 + n_2$ e $Y_i \sim iidN(\mu_3, \sigma^2)$ para $i = n_1 + n_2 + 1, \ldots, n$. Vemos que podemos formar tres grupos de observaciones en la muestra: el primero contiene las primeras n_1 observaciones; el segundo, las siguientes n_2 observaciones; y el tercero, las restantes n_3 observaciones. Queremos contrastar la hipótesis $H_0 : \mu_1 = \mu_2 = \mu_3$ versus $H_1 : H_0$ *es falsa* al nivel de significación α.

Ejemplo 13. *Queremos investigar la relación entre el salario medio de los trabajadores y la variable cualitativa nivel de estudios. Suponemos que las modalidades del nivel de estudios son: estudios obligatorios, estudios medios y estudios universitarios. Clasificamos las observaciones $Y_1, Y_2, \ldots, Y_n$ en tres grupos, y suponemos que las primeras n_1 corresponden a personas con estudios obligatorios; las siguientes n_2 observaciones, a personas con estudios medios; y las últimas n_3 observaciones, a personas con estudios universitarios. La hipótesis nula afirma que el salario medio es el mismo en los tres grupos.*

Para realizar el contraste consideramos la ecuación de regresión múltiple

$$Y_i = \mu_1 d_{1i} + \mu_2 d_{2i} + \mu_3 d_{3i} + u_i = 1, \quad i \ldots, n \tag{7.3}$$

en donde las variables binarias d_{1i}, d_{2i} y d_{3i} se definen del siguiente modo

$$d_{ji} = \begin{cases} 1 & \text{si } i \in \text{grupo } j \\ 0 & \text{si } i \notin \text{grupo } j \end{cases}$$

El estimador minimo-cuadrático del vector de parámetros (μ_1, μ_2, μ_3) es

$$\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n d_{1i}^2 & \sum_{i=1}^n d_{1i}d_{2i} & \sum_{i=1}^n d_{1i}d_{3i} \\ \sum_{i=1}^n d_{1i}d_{2i} & \sum_{i=1}^n d_{2i}^2 & \sum_{i=1}^n d_{2i}d_{3i} \\ \sum_{i=1}^n d_{1i}d_{3i} & \sum_{i=1}^n d_{2i}d_{3i} & \sum_{i=1}^n d_{3i}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n d_{1i}Y_i \\ \sum_{i=1}^n d_{2i}Y_i \\ \sum_{i=1}^n d_{3i}Y_i \end{pmatrix}$$

Teniendo en cuenta que hay n_1 observaciones en el primer grupo, n_2 en el segundo y n_3 en el tercero, y que cada observación pertenece a uno y sólo uno de los tres grupos tenemos que

$$\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{pmatrix} = \begin{pmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^n d_{1i}Y_i \\ \sum_{i=1}^n d_{2i}Y_i \\ \sum_{i=1}^n d_{3i}Y_i \end{pmatrix} = \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \bar{Y}_3 \end{pmatrix}$$

Vemos que la ecuación (7.3) nos permite estimar la media de la variable dependiente en cada uno de los tres grupos. Denotamos por SCR_1 la suma de cuadrados de este modelo.

Observación 38. *Si en la ecuación (7.3) se incluye un término constante, se incurre en la trampa de las variables ficticias.*

Bajo $H_0 : \mu_1 = \mu_2 = \mu_3$, tenemos el modelo restringido

$$Y_i = \mu + u_i \quad i = 1, \ldots, n$$

y denotamos su suma de cuadrados de los residuos por SCR_0. De modo que el contrate de la hipótesis de igualdad de medias puede basarse en el test de restricciones lineales formulado en términos de sumas de cuadrados

$$F = \frac{(SCR_0 - SCR_1)/2}{SCR_1/(n-3)} \sim F_{2,n-3}$$

La hipótesis $H_0 : \mu_1 = \mu_2 = \mu_3$ se rechaza al nivel de significación α, si $F > c$ en donde c es el valor crítico tal que $Prob(F_{2,n_3} > c) = \alpha$.

Dado que $d_{1i} + d_{2i} + d_{3i} = 1$, podemos reemplazar d_{1i} por $1 - d_{2i} - d_{3i}$ y reformular la ecuación 7.3 como

$$Y_i = \mu_1(1 - d_{2i} - d_{3i}) + \mu_2 d_{2i} + \mu_3 d_{3i} + u_i \quad i = 1, \ldots, n$$

o bien

$$Y_i = \mu_1 + (\mu_2 - \mu_1)d_{2i} + (\mu_3 - \mu_1)d_{3i} + u_i \quad i = 1, \ldots, n$$

que es la regresión de Y_i sobre un término constante, d_{2i} y d_{3i}

$$Y_i = \beta_1 + \beta_2 d_{2i} + \beta_3 d_{3i} + u_i \quad i = 1, \ldots, n$$

La variable dummy omitida es la correspondiente al grupo 1, que es el grupo base. La ordenada estimada $\hat{\beta}_1$ es la media de la variable dependiente para el grupo base $\hat{\mu}_1 = \bar{Y}_1$, y la pendiente estimada asociada a la dummy j, $\hat{\beta}_j$, es la diferencia entre la media del grupo específico j y la media del grupo base, $\hat{\mu}_j - \hat{\mu}_1 = \bar{Y}_j - \bar{Y}_1$, que se denomina efecto diferencial de la modalidad o factor j.

Aquí, la $H_0 : \mu_1 = \mu_2 = \mu_3$ es equivalente a $H_0 : \beta_2 = \beta_3$, que puede contrastarse con el test de significación global

$$F = \frac{R^2/2}{(1-R^2)/(n-3)} \sim F_{2,n-3}$$

La hipótesis $H_0 : \mu_1 = \mu_2 = \mu_3$ se rechaza al nivel de significación α, si $F > c$ en donde c es el valor crítico tal que $Prob(F_{2,n-3} > c) = \alpha$.

En resumen, para comparar las medias de m distribuciones normales podemos especificar las ecuaciones de regresión:

1. $Y_i = \alpha_1 d_{1i} + \cdots + \alpha_m d_{mi} + u_i$
2. $Y_i = \beta_1 + \beta_2 d_{2i} + \cdots + \beta_m d_{mi} + u_i$

en donde $\hat{\alpha}_j = \bar{Y}_j$ para $j = 1, \ldots, m$; $\hat{\beta}_1 = \bar{Y}_1$ y $\hat{\beta}_j = \bar{Y}_j - \bar{Y}_1$ para $j = 2, \ldots, m$.

7.2.3. Análisis de varianza de un sólo factor. El análisis que hemos desarrollado para comparar las medias de dos o más distribuciones normales se conoce como análisis de varianza de un sólo factor; en inglés, ANOVA one-way layout.

Suponemos que las observaciones Y_1, $Y_2, \ldots, Y_n$ provienen de m distribuciones normales con medias $\mu_j \, (j = 1, \ldots, m)$ y varianza común σ^2. Podemos, por tanto, formar m grupos de observaciones, teniendo cada grupo un tamaño muestral n_j y cumpliéndose que $n_1 + n_2 + \cdots + n_m = n$.

Es conveniente denotar por G_j el conjunto de observaciones pertenencientes al grupo j, $G_j = \{i : Y_i \sim N(\mu_j, \sigma^2)\}$. Así, podemos expresar la media y varianza de las observaciones pertenecientes a este grupo como

$$\bar{Y}_j = \frac{\sum_{i \in G_j} Y_i}{n_j} \quad \text{y} \quad s_j^2 = \frac{\sum_{i \in G_j} (Y_i - \bar{Y}_j)^2}{n_j}$$

y la media y varianza de todas las observaciones como

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{1}{n} \sum_{j=1}^m n_j \bar{Y}_j \quad \text{y} \quad s^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n} \sum_{j=1}^m \sum_{i \in G_j} (Y_i - \bar{Y})^2$$

Definición 49. *La suma de cuadrados total dentro del grupo j es*

$$SCT_j = \sum_{i \in G_j} (Y_i - \bar{Y}_j)^2$$

Definición 50. *La suma de cuadrados total intra-grupos (within) es*

$$SCT_w = \sum_{j=1}^m \sum_{i \in G_j} (Y_i - \bar{Y}_j)^2 = \sum_{j=1}^m SCT_j$$

Definición 51. *La suma de cuadrados total entre grupos (between) es*

$$SCT_b = \sum_{j=1}^m n_j (\bar{Y}_j - \bar{Y})^2$$

Proposición 62. *La suma de cuadrados total (SCT) puede particionarse en la suma de cuadrados intra-grupos (SCT_w) y la suma de cuadrados entre-grupos (SCT_b)*

$$SCT = SCT_w + SCT_b$$

DEMOSTRACIÓN.

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 &= \sum_{j=1}^{m}\sum_{i\in G_j}(Y_i-\bar{Y})^2 = \sum_{j=1}^{m}\sum_{i\in G_j}[(Y_i-\bar{Y}_j)+(\bar{Y}_j-\bar{Y})]^2 \\ &= \sum_{j=1}^{m}\sum_{i\in G_j}(Y_i-\bar{Y}_j)^2+\sum_{j=1}^{m}\sum_{i\in G_j}(\bar{Y}_j-\bar{Y})^2+2\sum_{j=1}^{m}\sum_{i\in G_j}(Y_i-\bar{Y}_j)(\bar{Y}_j-\bar{Y}) \\ &= \sum_{j=1}^{m}\sum_{i\in G_j}(Y_i-\bar{Y}_j)^2+\sum_{j=1}^{m}n_j(\bar{Y}_j-\bar{Y})^2\end{aligned}$$

en donde se ha usado el resultado $\sum_{i\in G_j}(Y_i-\bar{Y}_j)=0$ □

Queremos contrastar la hipótesis de que las m distribuciones tienen la misma media

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

$$H_1 : H_0 \text{ es falsa}$$

Podemos formular el contraste de igualdad de medias en el marco del modelo de regresión lineal. Así, bajo H_0 tenemos el modelo de regresión

$$Y_i=\mu+u_i$$

en donde $u_i\sim N(0,\sigma^2)$. El estimador mínimo-cuadrático de μ es $\hat{\mu}=\bar{Y}$ y la suma de cuadrados de los residuos $SCR_0=\sum_{i=1}^{n}(Y_i-\bar{Y})^2$.

Bajo H_1 tenemos el modelo de regresión

$$Y_i=\mu_1 d_{1i}+\cdots+\mu_m d_{mi}+u_i$$

en donde $u_i\sim N(0,\sigma^2)$. El estimador mínimo-cuadrático de μ_j es $\hat{\mu}_j=\bar{Y}_j$ y la suma de cuadrados de los residuos

$$SCR_1=\sum_{i=1}^{n}(Y_i-\bar{Y}_1 d_{1i}-\cdots-\bar{Y}_m d_{mi})^2$$

que puede escribirse como

$$SCR_1=\sum_{i\in G_1}(Y_i-\bar{Y}_1)^2+\cdots+\sum_{i\in G_m}(Y_i-\bar{Y}_m)^2=SCT_w$$

El estadístico de contrate para estas restricciones lineales es

$$F=\frac{SCR_0-SCR_1}{SCR_1}\frac{n-m}{m-1}=\frac{SCT_b/(m-1)}{SCT_w/(n-m)}\sim F_{m-1,n-m}$$

Este contraste suele presentarse en una tabla, denominada ANOVA, que tiene la siguiente forma

Fuente de variación	Grados de libertad	Suma de cuadrados	Media cuadrática
Entre-grupos	$m-1$	SCT_b	$SCT_b/(m-1)$
Intra-grupos	$n-m$	SCT_w	$SCT_w/(n-m)$
Total	$n-1$	SCT	

7.3. Modelo de regresión con varias variables cualitativas

Observe que en el análisis presentado, la variable cualitativa induce una clasificación de la muestra en dos o más grupos. Cada uno de los grupos está representado en la ecuación de regresión por una variable binaria, cuyo coeficiente estimado es la media de ese grupo. Al cambiar la variable binaria por una constante, la ordenada estimada es la media del grupo base, y las otras pendientes expresan las diferencias en los valores medios respecto al grupo base.

Consideramos ahora dos variables cualitativas D y F, con m y p modalidades: $D_{1i}, \ldots, D_{mi}$ y $F_{1i}, \ldots, F_{pi}$. Entonces ambas inducen una clasificación de las observaciones de la muestra en $m \times p$ grupos que puede mostrarse en la siguiente tabla

	F_1	F_2	$\ldots$	F_p
D_1	n_{11}	n_{12}	$\ldots$	n_{1p}
D_2	n_{21}	n_{22}	$\ldots$	n_{2p}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
D_m	n_{m1}	n_{m2}	$\ldots$	n_{mp}

EJEMPLO 14. *Los datos de salarios para n personas pueden clasificarse en términos de las modalidades de las variables cualitativas sexo y nivel de estudios en seis grupos*

	E. obligatorios	*E. medios*	*E. universitarios*
Mujeres	n_{11}	n_{12}	n_{13}
Hombres	n_{21}	n_{22}	n_{23}

Siguiendo un análisis similar al desarrollado para el modelo de regresión con una variable cualitativa, parece razonable especificar el modelo de regresión

$$Y_i = \alpha_1 D_{1i} + \cdots + \alpha_m D_{mi} + \beta_1 F_{1i} + \cdots + \beta_p F_{pi} + u_i \tag{7.4}$$

Sin embargo, este modelo presenta multicolinealidad exacta porque la suma de las variables binarias asociadas a la primera variable cualitativa es igual a la suma de las variables binarias asociadas a la segunda variable cualitativa. De manera que, las variables explicativas son linealmente dependientes. Surge aquí otra forma de **la trampa de las variables ficticias**.

Los $m \times p$ grupos inducidos por las variables cualitativas D y F pueden recogerse en la siguiente ecuación de regresión

$$Y_i = \alpha + \beta_2 D_{2i} + \cdots + \beta_m D_{mi} + \delta_2 F_{2i} + \cdots + \delta_p F_{pi} + u_i \tag{7.5}$$

Comparando (7.4) y (7.5), vemos que hemos omitido una variable una variable dummy por cada variable cualitativa y hemos incluido un término constante. Las modalidades omitidas determinan el grupo base respecto del que se realizan las comparaciones.

En (7.5), el valor esperado de la variable dependiente es

$$E(Y_i) = \begin{cases} \alpha & i \in Grupo(1,1) \\ \alpha + \beta_j + \delta_h & i \in Grupo(j,h) \text{ para } j = 2, \ldots, m;\ h = 2, \ldots, p \end{cases}$$

Ejemplo 15. *En la regresión del salario sobre el sexo y el nivel de estudios*

$$Y_i = \alpha + \beta_2 H_i + \delta_2 EM_i + \delta_3 EU_i + u_i$$

el grupo base es el de mujeres con estudios obligatorios. El salario esperado para el grupo base es α; para el grupo de hombres con estudios obligatorios, $\alpha + \beta_2$; para mujeres con estudios medios, $\alpha + \delta_2$; para hombres con estudios medios, $\alpha + \beta_2 + \delta_2$; para mujeres con estudios universitarios, $\alpha + \delta_3$; y para hombres con estudios universitarios, $\alpha + \beta_2 + \delta_3$.

	E. obligatorios	*E. medios*	*E. universitarios*
Mujeres	α	$\alpha + \delta_2$	$\alpha + \delta_3$
Hombres	$\alpha + \beta_2$	$\alpha + \beta_2 + \delta_2$	$\alpha + \beta_2 + \delta_3$

Cuadro 1: Salario esperado por sexo y nivel de estudios

Las hipótesis que nos interesa contrastar son $H_0 : \beta_2 = \cdots = \beta_m = 0$ y $H_0 : \delta_2 = \cdots = \delta_p = 0$, que son hipótesis de significación conjunta de un subconjunto de coeficientes de regresión. El contraste de este tipo de hipótesis es el objetivo del análisis de varianza de dos factores; en inglés, ANOVA two-way layout.

Ejercicio 3. *Especifique un modelo de regresión que explique la deducción por vivienda en el IRPF, Y_i, en función de la actividad profesional (agricultor, trabajador por cuenta ajena, trabajador por cuenta propia, rentista) y del lugar de residencia (rural, urbano) del declarante. ¿Cómo contrastaría la hipótesis de que esta deducción beneficia a los rentistas urbanos?*

7.3.1. Efectos de interacción. Una posible limitación de la ecuación de regresión

$$Y_i = \alpha + \beta_2 H_i + \delta_2 EM_i + \delta_3 EU_i + u_i$$

es que la diferencia en el valor esperado de Y_i para una mujer y un hombre es β_2, con independencia de su nivel de estudios. Del mismo modo, la diferencia en el valor esperado de Y_i para un universitario y una persona con estudios obligatorios es β_3, con independencia de si es hombre o mujer.

Una forma de eliminar esta deficiencia consiste en incluir los denominados efectos de interacción

$$Y_i = \alpha + \beta_2 H_i + \delta_2 EM_i + \delta_3 EU_i + \gamma_2 H_i EM_i + \gamma_3 H_i EU_i + u_i$$

Ahora, la tabla de valores esperados de Y_i para los distintos grupos en la muestra es

	E. obligatorios	E. medios	E. universitarios
Mujeres	α	$\alpha + \delta_2$	$\alpha + \delta_3$
Hombres	$\alpha + \beta_2$	$\alpha + \beta_2 + \delta_2 + \gamma_2$	$\alpha + \beta_2 + \delta_3 + \gamma_3$

donde vemos que la diferencia en el valor esperado de Y_i para una mujer y un hombre es: β_2, si ambos tienen estudios obligatorios; $\beta_2 + \gamma_2$, si ambos son bachilleres; y $\beta_2 + \gamma_3$, si ambos son universitarios.

7.4. Cambio estructural: test de Chow

Otra de las aplicaciones de las variables ficticias es el denominado contraste de cambio estructural.

DEFINICIÓN 52. *Se habla de cambio estructural o ruptura estructural cuando los parámetros de la ecuación de regresión difieren en las distintas submuestras.*

La posibilidad de un cambio estructural en la estimación de modelos econométricos con datos de series temporales debe tenerse en cuenta cuando en la muestra acontecen sucesos tales como

1. la existencia de periodos de paz y de guerra
2. cambios de metodología en la elaboración de los datos (el paso del SCN-1968 y SEC-1979 al SCN-1993 y SEC-95 en las cuentas nacionales)
3. cambios políticos (dictadura-democracia)
4. cambios legislativos, etc.

En el caso de datos de sección cruzada, un ejemplo de cambio estructural es la discriminación sexual en el salario. Imaginemos que disponemos de datos sobre el salario, Y, y la experiencia laboral en años, X, para una muestra de empleados en una empresa multinacional. Si especificamos la ecuación de regresión simple

$$Y_i = \alpha + \beta X_i + u_i \quad i = 1, \ldots, n$$

entonces, suponiendo que $E(u_i) = 0$, el salario esperado para un trabajador i con una experiencia laboral X_i es

$$E(Y_i) = \alpha + \beta X_i$$

La ordenada α se interpreta como el salario inicial esperado para una persona sin experiencia laboral, y la pendiente β como el aumento salarial esperado por cada año de experiencia laboral.

Esta ecuación de regresión, junto con las variables ficticias, nos permite expresar tres formas diferentes de discriminación sexual en el salario:

1. salarios iniciales diferentes y aumentos salariales iguales (ordenadas diferentes y pendientes iguales),
2. salarios iniciales iguales y aumentos salariales diferentes (ordenadas iguales y pendientes diferentes),
3. salarios iniciales y aumentos salariales diferentes (ordenadas y pendientes diferentes).

La primera forma de discriminación sexual puede expresarse mediante la ecuación de regresión

$$Y_i = \alpha_M M_i + \alpha_H H_i + \beta X_i + u_i \quad i = 1, \ldots, n$$

donde vemos que el salario esperado para una mujer es

$$E(Y_i) = \alpha_M + \beta X_i$$

y el salario esperado para un hombre

$$E(Y_i) = \alpha_H + \beta X_i$$

Observamos que, en las dos submuestras de mujeres y hombres, las ordenadas (salarios iniciales) son diferentes y las pendientes (aumentos salariales) son iguales. Esta forma de discriminación sexual puede contrastarse calculando el estadístico t para la hipótesis nula $H_0 : \alpha_M = \alpha_H$ frente a la alternativa $H_1 : \alpha_M \neq \alpha_H$

$$t = \frac{\hat{\alpha}_M - \hat{\alpha}_H}{\sqrt{\hat{V}(\hat{\alpha}_M) + \hat{V}(\hat{\alpha}_H)}} \sim t_{n-3}$$

De forma equivalente, podemos escribir la primera forma de discriminación

$$Y_i = \gamma_1 + \gamma_2 H_i + \gamma_3 X_i + u_i \quad i = 1, \ldots, n$$

donde $\gamma_1 = \alpha_M$, $\gamma_2 = \alpha_H - \alpha_M$ y $\gamma_3 = \beta$. El salario esperado para una mujer es

$$E(Y_i) = \gamma_1 + \gamma_3 X_i$$

y el salario esperado para un hombre

$$E(Y_i) = \gamma_1 + \gamma_2 + \gamma_3 X_i$$

recogiendo γ_2 la diferencia en los salarios iniciales de mujeres y hombres. El contraste de la hipótesis de que no hay discriminación sexual puede basarse en el contraste de significación individual de γ_2.

Análogamente, la segunda forma de discriminación salarial corresponde a

$$Y_i = \alpha + \beta_M X_i M_i + \beta_H X_i H_i + u_i \quad i = 1, \ldots, n$$

donde vemos que el salario esperado para una mujer es

$$E(Y_i) = \alpha + \beta_M X_i$$

y el salario esperado para un hombre

$$E(Y_i) = \alpha + \beta_H X_i$$

Ahora las ordenadas son iguales, pero las pendientes son diferentes. Esta forma de discriminación sexual puede contrastarse calculando el estadístico t para la hipótesis nula $H_0 : \beta_M = \beta_H$ frente a la alternativa $H_1 : \beta_M \neq \beta_H$

$$t = \frac{\hat{\beta}_M - \hat{\beta}_H}{\sqrt{\hat{V}(\hat{\beta}_M) + \hat{V}(\hat{\beta}_H)}} \sim t_{n-3}$$

La representación equivalente usando una variable ficticia es

$$Y_i = \gamma_1 + \gamma_2 X_i + \gamma_3 X_i H_i + u_i \quad i = 1, \ldots, n$$

donde $\gamma_1 = \alpha$, $\gamma_2 = \beta_M$ y $\gamma_3 = \beta_H - \beta_M$. El salario esperado para una mujer es

$$E(Y_i) = \gamma_1 + \gamma_2 X_i$$

y el salario esperado para un hombre

$$E(Y_i) = \gamma_1 + (\gamma_2 + \gamma_3) X_i$$

recogiendo γ_3 la diferencia en los aumentos salariales de mujeres y hombres.

Por último, la tercera forma de discriminación salarial sería

$$Y_i = \alpha_M M_i + \alpha_H H_i + \beta_M X_i M_i + \beta_H X_i H_i + u_i \quad i = 1, \ldots, n$$

donde el salario esperado para una mujer es

$$E(Y_i) = \alpha_M + \beta_M X_i$$

y el salario esperado para un hombre

$$E(Y_i) = \alpha_H + \beta_H X_i$$

Tanto las ordenadas como las pendientes son diferentes en las dos submuestras de mujeres y hombres. La hipótesis de no discriminación sexual, $H_0 : \alpha_M = \alpha_H$, $\beta_M = \beta_H$, puede basarse en un contraste F de restricciones lineales con

$$\mathbf{R} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha_M \\ \alpha_H \\ \beta_M \\ \beta_H \end{pmatrix} \quad \mathbf{r} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

La representación equivalente usando una variable ficticia sería

$$Y_i = \gamma_1 + \gamma_2 H_i + \gamma_3 X_i + \gamma_4 X_i H_i + u_i \quad i = 1, \ldots, n$$

donde $\gamma_1 = \alpha_M$, $\gamma_2 = \alpha_M - \alpha_H$, $\gamma_3 = \beta_M$, $\gamma_4 = \beta_M - \beta_H$. La hipótesis de no discriminación sexual, $H_0 : \gamma_2 = 0$, $\gamma_4 = 0$, puede basarse en un contraste F de restricciones lineales con

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix} \quad \mathbf{r} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

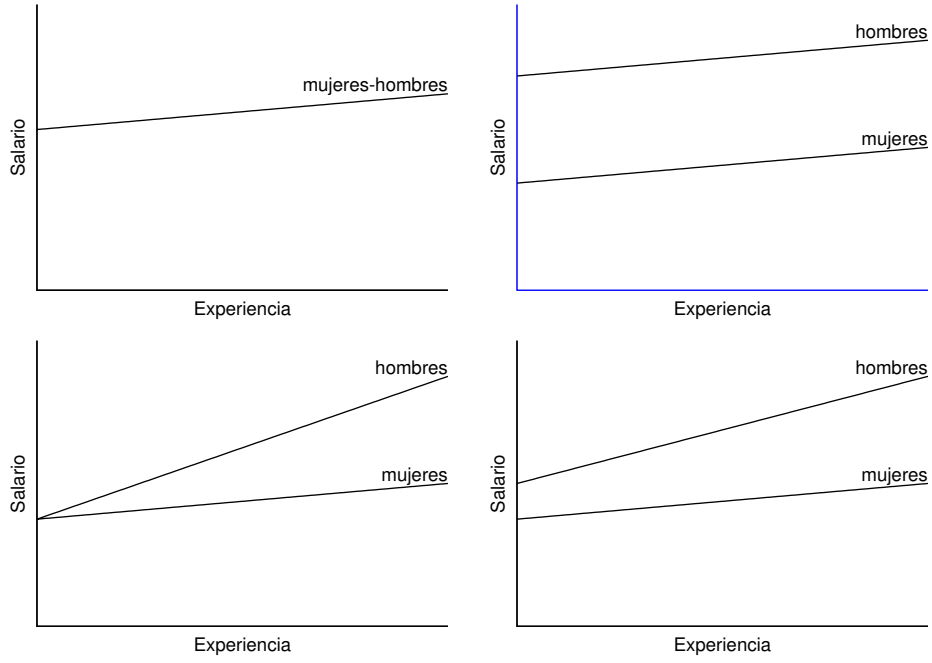

Figura 1: Cambio estructural en regresión simple

La figura 1 ilustra el caso de no discriminación sexual y las tres posibles formas de discriminación sexual estudiadas. Si contemplamos la figura 1 como una matriz de gráficos, entonces el gráfico (1,1) describe el caso de no discriminación salarial, ordenadas y pendientes iguales. El gráfico (1,2) corresponde a ordenadas diferentes y pendientes iguales, y muestra que, en cada nivel de experiencia, las mujeres tienen un salario esperado menor que los hombres, siendo estas diferencias salariales constantes. El gráfico (2,1) corresponde a ordenadas iguales y pendientes diferentes, y muestra que el salario inicial esperado es el mismo para mujeres y hombres, pero que las diferencias salariales esperadas aumentan con la experiencia. Finalmente, el gráfico (2,2) corresponde a ordenadas y pendientes diferentes, y revela que la discriminación salarial existe en todos los niveles de experiencia y que aumenta con ésta.

La anterior ilustración del cambio estructural se extiende fácilmente al modelo lineal general

$$\mathbf{y} = \mathbf{i}\alpha + \mathbf{X}_s\boldsymbol{\beta}_s + \mathbf{u}$$

Si consideramos dos submuestras, podemos particionar los datos del siguiente modo

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} \mathbf{i} & \mathbf{X}_s \end{pmatrix} = \begin{pmatrix} \mathbf{i}_1 & \mathbf{X}_{s1} \\ \mathbf{i}_2 & \mathbf{X}_{s2} \end{pmatrix}$$

donde $\mathbf{y}_1$ es el vector $n_1 \times 1$ de observaciones de la variable dependiente en la primera submuestra, $\mathbf{i}_1$ es un vectos $n_1 \times 1$ de unos y $\mathbf{X}_{s1}$ es la matriz $n_1 \times (k-1)$ que contiene los datos de las $k-1$ variables explicativas en la primera submuestra. Análogamente, se definen $\mathbf{y_2}$, $\mathbf{i}_2$ y $\mathbf{X}_{s2}$. Note que donde el subíndice 1 indica la primera submuestra y el subíndice 2 la segunda. Dada esta partición de los datos, el modelo lineal general puede escribirse como

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{i}_1 & \mathbf{X}_{s1} \\ \mathbf{i}_2 & \mathbf{X}_{s2} \end{pmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\beta}_s \end{pmatrix} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} \tag{7.6}$$

En esta especificación podemos considerar tres tipos de cambio estructural:

1. Ordenadas diferentes y pendientes iguales

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{i}_1 & \mathbf{0}_1 & \mathbf{X}_1 \\ \mathbf{0}_2 & \mathbf{i}_2 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \boldsymbol{\beta}_s \end{pmatrix} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} \tag{7.7}$$

   donde $\mathbf{0}_1$ y $\mathbf{0}_2$ son vectores $n_1 \times 1$ y $n_2 \times 1$ de ceros.

2. Ordenadas iguales y pendientes diferentes

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{i}_1 & \mathbf{X}_1 & \mathbf{O}_1 \\ \mathbf{i}_2 & \mathbf{O}_2 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \boldsymbol{\beta}_{s1} \\ \boldsymbol{\beta}_{s2} \end{pmatrix} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} \tag{7.8}$$

   donde $\mathbf{O}_1$ y $\mathbf{O}_2$ son matrices $n_1 \times (k-1)$ y $n_2 \times (k-1)$ de ceros.

3. Ordenadas y pendientes diferentes

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{i}_1 & \mathbf{0}_1 & \mathbf{X}_1 & \mathbf{O}_1 \\ \mathbf{0}_2 & \mathbf{i}_2 & \mathbf{O}_2 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \boldsymbol{\beta}_{s1} \\ \boldsymbol{\beta}_{s2} \end{pmatrix} + \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} \tag{7.9}$$

El test de Chow es un contraste de cambio estructural basado en el estadístico F de sumas de cuadrados de residuos

$$F = \frac{(SCR_0 - SCR_1)/(GL_0 - GL_1)}{SCR_1/GL_1} \sim F_{GL_0-GL_1,GL_1}$$

donde SCR_0 es la suma de cuadrados de los residuos bajo la hipótesis nula o suma de cuadrados del modelo restringido, GL_0 son los grados de libertad del modelo restringido, SCR_1 es la suma de cuadrados de los residuos bajo la hipótesis alternativa o suma de cuadrados de los residuos en el modelo sin restricciones, y GL_1 son los grados de libertad en el modelo sin restricciones.

Los pasos para realizar el contraste de homogeneidad de ordenadas

1. Estimar el modelo (7.7) bajo $H_0 : \alpha_1 = \alpha_2$, que se convierte en (7.6), y calcular la suma de cuadrados de los residuos SCR_0 y los grados de libertad $GL_0 = n-k$.
2. Estimar el modelo (7.7) bajo $H_1 : \alpha_1 \neq \alpha_2$, calcular la suma de cuadrados de los residuos, SCR_1, y los grados de libertad $GL_1 = n - k - 1$.
3. Calcular el estadístico de contraste

$$F = \frac{(SCR_0 - SCR_1)/1}{SCR_1/(n-k-1)} \sim F_{1,n-k-1}$$

En el contraste de homogeneidad de pendientes seguimos los siguientes pasos

1. Estimar el modelo (7.8) bajo $H_0 : \boldsymbol{\beta}_{s1} = \boldsymbol{\beta}_{s2}$, que se convierte en (7.6), y calcular la suma de cuadrados de los residuos SCR_0 y los grados de libertad $GL_0 = n - k$.
2. Estimar el modelo (7.8) bajo $H_1 : \boldsymbol{\beta}_{s1} \neq \boldsymbol{\beta}_{s2}$, calcular la suma de cuadrados de los residuos, SCR_1, y los grados de libertad $GL_1 = n - 2k - 1$.
3. Calcular el estadístico de contraste

$$F = \frac{(SCR_0 - SCR_1)/(k-1)}{SCR_1/(n-2k-1)} \sim F_{k-1,n-2k-1}$$

Finalmente, en el contraste de homogeneidad de ordenadas y pendientes seguimos los siguientes pasos

1. Estimar el modelo (7.9) bajo $H_0 : \alpha_1 = \alpha_2, \boldsymbol{\beta}_{s1} = \boldsymbol{\beta}_{s2}$, que se convierte en (7.6), y calcular la suma de cuadrados de los residuos SCR_0 y los grados de libertad $GL_0 = n - k$.
2. Estimar el modelo (7.8) bajo $H_1 : \alpha_1 \neq \alpha_2, \boldsymbol{\beta}_{s1} \neq \boldsymbol{\beta}_{s2}$, calcular la suma de cuadrados de los residuos, SCR_1, y los grados de libertad $GL_1 = n - 2k$.
3. Calcular el estadístico de contraste

$$F = \frac{(SCR_0 - SCR_1)/k}{SCR_1/(n-2k)} \sim F_{k,n-2k}$$

Ejercicio 4. *Extienda el test de Chow al caso de tres submuestras.*

7.5. Predicción de series temporales

El gráfico temporal en la figura 2 muestra la evolución de la serie mensual de ingresos por turismo en España durante el periodo muestral comprendido entre enero de 1990 y abril de 2007. La serie presenta dos características estadísticas muy obvias: crecimiento lineal y estacionalidad (comportamiento periódico que se repite todos los años). Estas dos características o **hechos estilizados** aparecen en un buen número de series mensuales

y trimestrales, y pueden ser descritas usando un modelo de regresión con variables explicativas deterministas dependientes del tiempo.

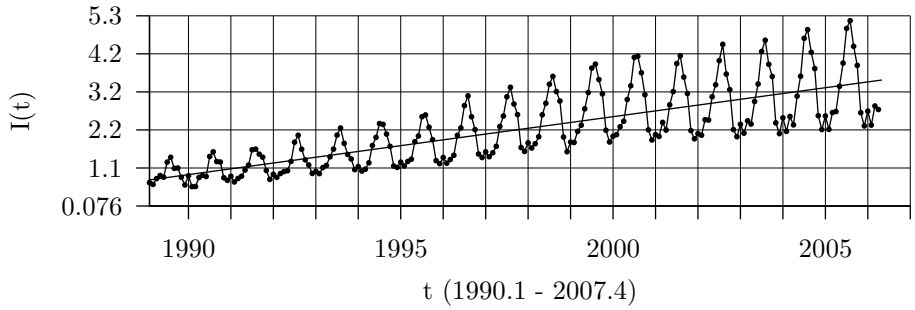

Figura 2: Ingresos por turismo en España (millones de euros)

7.5.1. Tendencia determinista. El modelo de regresión con una tendencia lineal es

$$y_t = \beta_0 + \beta_1 t + u_t, \quad t = 1, \ldots, n$$

en donde t es el índice que ordena los datos y la variable explicativa. La ordenada β_0 es del valor esperado de y_t para $t = 0$, y la pendiente indica la variación esperada en la variable dependiente entre dos instantes temporales consecutivos, $\beta_1 = E(y_t - y_{t-1})$. Las estimaciones de mínimos cuadrados de estos dos parámetros son

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n}(t - \frac{n+1}{2})y_t}{\sum_{t=1}^{n}(t - \frac{n+1}{2})^2} = \frac{\sum_{t=1}^{n}(t - \frac{n+1}{2})y_t}{n^3 - n} \quad \text{y} \quad \hat{\beta}_0 = \bar{y} - \beta_1 \frac{n+1}{2}$$

El modelo de tendencia lineal es un caso especial del modelo con tendencia polinomial de orden r

$$y_t = \beta_0 + \beta_1 t + \cdots + \beta_r t^r + u_t, \quad t = 1, \ldots, n \tag{7.10}$$

7.5.2. Estacionalidad determinista. La estacionalidad presente en una serie trimestral puede describirse mediante una combinación lineal de 4 variables ficticias estacionales

$$y_t = \alpha_1 d_{1t} + \alpha_2 d_{2t} + \alpha_3 d_{3t} + \alpha_4 d_{4t} + u_t, \quad t = 1, \ldots, n \tag{7.11}$$

en donde $d_{jt}\,(j = 1, \ldots, 4)$ toma el valor 1 si la observación t-ésima corresponde al trimestre j, y 0 en cualquier otro caso

$$d_{jt} = \begin{cases} 1 & t \in Trimestre\, j \\ 0 & t \notin Trimestre\, j \end{cases}$$

Es conveniente notar que las variables ficticias estacionales son mutuamente ortogonales: si la observación t-ésima corresponde al primer trimestre $d_{1t} = 1$ y $d_{2t} = d_{3t} = d_{4t} = 0$. De aquí, los coeficientes de regresión $\alpha_j\,(j = 1, \ldots, 4)$ pueden estimarse fácilmente en las regresiones

$$y_t = \alpha_j d_{jt} + v_t \quad t = 1, \ldots, n$$

siendo

$$\hat{\alpha}_j = \frac{\sum_{t=1}^{n} y_t d_{jt}}{\sum_{t=1}^{n} d_{jt}^2} = \frac{y_j + y_{j+4} + \cdots + y_{n-j+1}}{n/4} = \bar{y}_j$$

la media muestral de todas las observaciones del trimestre j. El cuadro 2, denominado tabla de Buys-Ballot, presenta los datos de una serie trimestral en una tabla de doble entrada cuyas filas indican el año y cuyas columnas indican el trimestre. En esta tabla podemos ver que $\sum_{t=1}^{n} y_t d_{jt}$ se corresponde con la suma de los datos de la columna j-ésima.

Año	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
1	y_1	y_2	y_3	y_4
2	y_5	y_6	y_7	y_8
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n/4	y_{n-3}	y_{n-2}	y_{n-1}	y_n

Cuadro 2: Tabla Buys-Ballot para una serie trimestral

En el caso de una serie mensual (doce datos por año), definiendo $d_{jt} = 1\,(j = 1, \ldots, 12)$ si la observación t corresponde al mes j, y $d_{jt} = 0$ en otro caso, tenemos

$$y_t = \beta_1 d_{1t} + \beta_2 d_{2t} + \cdots + \beta_{12} d_{12t} + u_t, \quad t = 1, \ldots, n$$

El coeficiente estimado $\hat{\beta}_j$ es la media de los datos correspondientes al mes j, para $j = 1, \ldots, 12$. Como $\sum_{j=1}^{12} d_{jt} = 1$, podemos reescribir la ecuación como

$$y_t = \alpha_1 + \alpha_2 d_{2t} + \cdots + \alpha_{12} d_{12t} + u_t, \quad t = 1, \ldots, n$$

en donde el mes base corresponde a enero. Ahora $\hat{\alpha}_1 = \hat{\beta}_1$ es la media de los datos correspondientes a enero y $\hat{\alpha}_j = \hat{\beta}_j - \hat{\beta}_1$ es la diferencia entre la media de los datos correspondientes al mes j y la media de los datos correspondientes al mes de enero. También podemos especificar la ecuación

$$y_t = \delta_0 + \delta_1 d_{1t} + \delta_2 d_{2t} + \cdots + \delta_{12} d_{12t} + u_t, \quad t = 1, \ldots, n$$

y evitar la trampa de las variables ficticias imponiendo la restricción $\delta_1 + \delta_2 + \cdots + \delta_{12} = 0$ que conduce al modelo

$$y_t = \beta_0 + \beta_2(d_{2t} - d_{1t}) + \cdots + \beta_{12}(d_{12t} - d_{1t}) + u_t, \quad t = 1, \ldots, n$$

en donde $\hat{\beta}_0$ es la media muestral de Y, y $\hat{\beta}_j$ es la diferencia entre la media de Y en el mes j menos la media global $\bar{Y}$.

En general, la estacionalidad de una serie temporal con periodo estacional s (número de observaciones por año) puede describirse de tres formas equivales

$$\begin{aligned} y_t &= \sum_{j=1}^{s} \alpha_j d_{jt} + u_t, \quad t = 1, \ldots, n \\ y_t &= \beta_1 + \sum_{j=2}^{s} \beta_j d_{jt} + u_t, \quad t = 1, \ldots, n \\ y_t &= \delta_0 + \sum_{j=2}^{s} \delta_j (d_{jt} - d_{1t}) + u_t, \quad t = 1, \ldots, n \end{aligned} \tag{7.12}$$

en donde

$$d_{jt} = \begin{cases} 1 & t \in \text{estación } j \\ 0 & t \notin \text{estación } j \end{cases}$$

7.6. Tendencia y estacionalidad determinista

Combinando los modelos (7.10)-(7.12) parece razonable especificar la relación

$$y_t = \beta_0 + \sum_{i=1}^{r} \beta_i t^i + \sum_{j=1}^{s} \alpha_j d_{jt} + u_t, \quad t = 1, \ldots, n \tag{7.13}$$

Sin embargo, esta ecuación presenta multicolinealidad exacta: la variable de unos asociada al termino constante es igual a la suma de las variables ficticias estacionales. Este problema, denominado la trampa de las variables ficticias, puede evitarse de tres formas:

1. omitiendo el término constante,
2. omitiendo una ficticia estacional,
3. imponiendo la restricción lineal $\alpha_1 + \cdots + \alpha_s = 0$.

Conviene notar que las predicciones de los valores futuros $y_t\,(t = n+1, \ldots, n+m)$ basadas en el modelo (7.13) son predicciones incondicionales porque las variables explicativas son deterministas, es decir, sus valores son conocidos en cualquier instante del tiempo.

Denotamos la predicción del valor futuro y_{n+h} basada en las observaciones disponibles hasta el instante n como $\hat{y}_n(h)$, en donde n indica el origen de predicción y h es el horizonte de predicción. Análogamente, el error de predicción en el origen n y al horizonte h se denota por $e_n(h)$. Puede comprobarse que los residuos mínimo cuadráticos $\hat{u}_t$ son errores de predicción $e_{t-1}(1)$.

El modelo de regresión con variables ficticias estacionales se utiliza para extraer o eliminar la estacionalidad de las series temporales trimestrales y mensuales. Los residuos resultantes de la estimación de este modelo proporcionan una serie corregida de variación estacional o serie desestacionalizada que muestra más claramente la evolución a largo plazo de la variable de interés.

Resumen

1. Una variable es determinista si sus valores son funciones exactas del índice observacional.
2. Las variables ficticias nos permiten comparar las medias de dos o más poblaciones.
3. El test de Chow es una aplicación de las variables ficticias para contrastar la estabilidad paramétrica.
4. El modelo de regresión con tendencia lineal y estacionalidad determinista se utiliza para predecir datos de series temporales y para eliminar la estacionalidad de las series mensuales y trimestrales.

Palabras clave

Variables cualitativas
Variables ficticias
La trampa de las variables ficticias
Cambio estructural
Variables ficticias estacionales
Tendencia lineal
Predicción de series temporales
Ajuste estacional