Regresión

Teresa Villagarcía

1. Introducción.

En cualquier ciencia es necesario establecer relaciones entre variables que se presupone que tienen relación con el objeto de estudio. Si estas relaciones existen y son comprendidas, podemos entender el funcionamiento del sistema y, consecuentemente, prever su comportamiento futuro, incluso si las condiciones varían. Existen dos tipos de relaciones claras entre variables:

1. Relaciones deterministas.
2. Relaciones no deterministas

Entendemos que una relación es determinista si podemos expresar la variable y en función de x :

$$y = f(x)$$

Esta relación indica que si $x = 20, y = f(20)$ siempre. Este modelo no permite ningún error de predicción y en la práctica no existe, pues hay errores en los aparatos de medida, que hacen que cuando $x = 20$ el valor de y varíe en torno a $f(20)$ de una medida a otra. Sin embargo en algunas circunstancias puede suponerse un modelo determinista. Por ejemplo un circuito eléctrico compuesto por una alimentación de 10 Voltios conectada a una resistencia de 5 Ohmios, tendrá una intensidad $I = \frac{V}{R} = \frac{10}{5} = 2$Amperios, y el error será despreciable, por lo que si conectamos el circuito una y otra vez obtendremos 2 Amperios de intensidad.

Muchas veces las relaciones son deconocidas o simplemente hay múltiples causas no asignables que hacen que el error en la medida sea muy grande. Así, cuando $x = 20$, obtenemos muchos valores de y. En este caso podemos pensar que cuando $x = 20, y \sim N(f(20), \sigma^2)$. Esto equivale a pensar que el modelo que siguen las variables es no determinista o probabilístico y puede formularse como

$$y = f(x) + u$$

donde u es una perturbación aleatoria.

2. Modelos lineales.

Los modelos lineales probabilísticos pueden expresarse de la siguiente forma

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

es decir el valor de y va variando linealmente con x, pero además existe un error aleatorio que hace que dos observaciones con la misma x no tengan necesariamente la misma y.

Ejemplo 1: Pesos y Alturas.

Los datos de la tabla 1 son los pesos en Kg y Alturas en cm de parte de un conjunto de 117 alumnos de la Universidad Politécnica de Madrid.

Altura	**Peso**
184	112
156	54
173	60
186	66
⋮	⋮
180	76

La Figura 1 presenta datos de Pesos y Alturas del Ejemplo 1. Como puede observarse, aunque no existe una relación exacta entre el Peso y la altura de los individuos, el conocer la altura de una persona ofrece importante información sobre su peso. Así, las personas más altas son también las que tienen más peso.

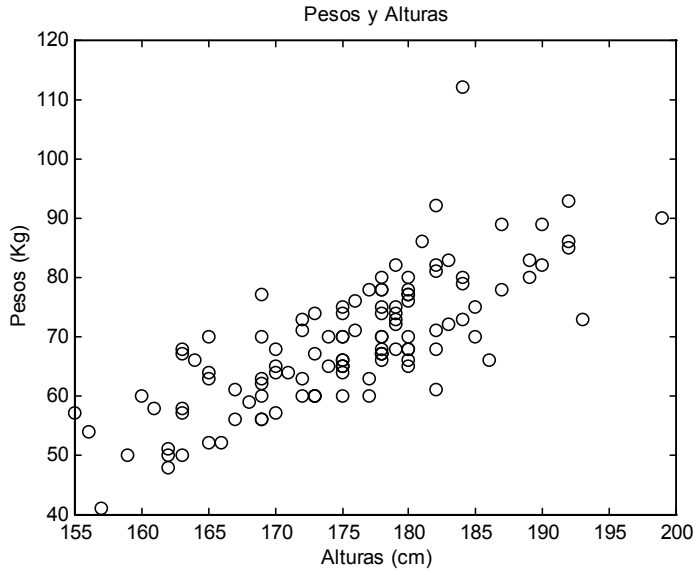


Figura 1: Pesos y Alturas de estudiantes de la Universidad

El análisis de regresión pretende encontrar una relación entre las variables Peso y Altura.

Para ello consideremos y la variable dependiente, o variable a explicar. La variable independiente o explicativa se denominará x . El modelo de regresión asume que los valores de la variable y pueden dividirse en dos partes: Una parte lineal explicada por la variable x , y una parte no explicada. La formulación del modelo es

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{2}$$

Esta relación indica que y varía linealmente con x , pero que no todo el valor de y se explica con x . En términos de nuestro ejemplo podemos decir que el peso de un individuo aumenta

linealmente a medida que aumenta su altura, pero además de la altura existen otros muchos factores determinantes del mismo. En la ecuación (2) la parte $\beta_0 + \beta_1 x_i$ representa la parte lineal de la variación de y respecto a x . Esta parte se denomina también parte determinista. La segunda parte, u_i, se denomina parte aleatoria y representa la parte de y no explicada linealmente por x .

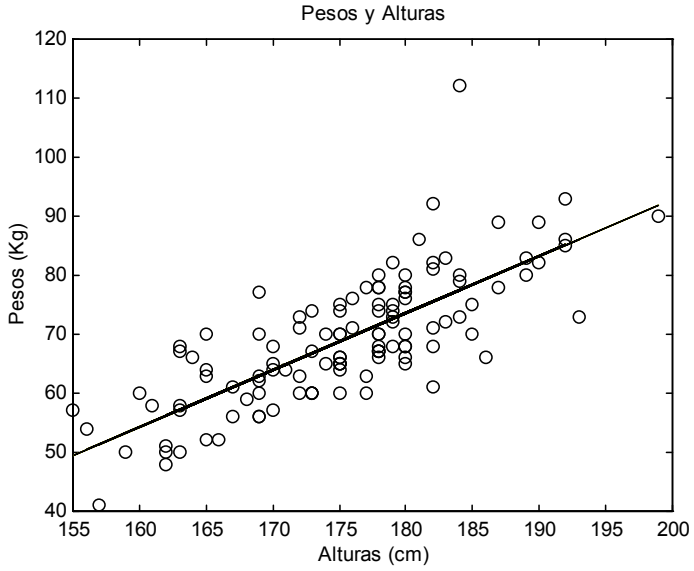


Figura 2: Datos de pesos y alturas con la recta de regresión.

La Figura 2 presenta la recta de regresión entre y y x . La recta ajustada tiene por ecuación

$$\text{Peso} = -100,22 + 0.97\text{Altura}$$

Así, una persona que mida 1,80 metros pesará según la recta de regresión $\hat{y} = -100,22 + 0,97\text{x}180 = 74,38\text{Kg}$. Indudablemente, no todas las personas que miden 1,80m pesan 74,38Kg. La diferencia entre el peso de una persona, y_i, y lo que la recta de regresión le predice, $\hat{y}_i$, se denomina resíduo.

Así, el último individuo de la tabla mide 1,80m y pesa 76Kg. El modelo estimado le predice un peso 74,38Kg. El resíduo o error en la predicción del peso es de 76-74,38. Este valor corresponde a la distancia vertical entre cada observación y la recta de regresión.

El residuo puede expresarse como $e_i = y_i - \hat{y}_i$.

2.1 Hipótesis básicas del modelo.

El modelo de regresión simple requiere hacer las siguientes hipótesis:

1. **Linealidad**: Es una hipótesis fundamental. Para ajustar una línea recta a un conjunto de datos es preciso que éstos tengan un aspecto razonablemente recto. Los datos del Ejemplo 1 cumplen esta hipótesis. La Figura 5 presenta un conjunto de datos cuya relación no es lineal.
2. $\mathbf{E}(\mathbf{u}_i) = 0$: El valor promedio del error es cero. Esta hipótesis implica que el ajuste se va realizar está centrado respecto a los datos. Así, en virtud de esta hipótesis cabe esperar que la

recta de regresión esté centrada en la nube de puntos de los datos.

3. $\mathbf{var}(\mathbf{u}_i) = \boldsymbol{\sigma}^2$ =cte: La varianza de los errores es constante y no depende del nivel de las variables. Esta hipótesis implica que la nube de puntos de los datos tiene una anchura semejante a lo largo de la recta de regresión. Si lo datos cumplen esta hipótesis se dice que son **Homocedástico**s. Por el contrario, datos cuya variablidad no es constante se denominana **Heterocedásticos**. La heterocedasticidad es algo muy corriente en numerosos fenómenos: Así, por ejemplo en economía cabe esperar que cuanto mayor sea el nivel de ingresos, mayor será la variabilidad en el gasto efectuado en consumo a lo largo del año para diversas unidades familiares.
4. **Independencia** $\mathbf{E}(\mathbf{u}_i\mathbf{u}_j) = 0$ Esta hipótesis implica que las observaciones son independientes. Es decir una observación (Peso, Altura) no ofrece información sobre los valores de la siguiente. Esta hipótesis invalida el análisis de regresión, al menos en sus versiones más simples, con datos de tipo temporal en los que exista autocorrelación. No es por tanto adecuado el análisis de regresión simple para estudiar cómo se relaciona la inflación en España con los tipos de interés a través del tiempo.
5. **Normalidad** $u_i \sim N(0, \sigma^2)$ Esta hipótesis se refiere a que los errores se distribuyen normalmente, es decir siguiendo una campana de Gauss. Es una hipótesis razonable en virtud del Teorema Central del Límite que dice que cuando una variable (Peso) es el resultado de muchas cosas pequeñas tenderá a distribuirse normalmente.

Es muy importante que se cumplan estas hipótesis. *Para ello se realizará un examen gráfico de los datos antes del ajuste y*, *posteriormente*, *se comprobará que se cumplen mediante el análisis de los residuos*.

2.2 Ajuste.

El ajuste del modelo se realiza por mínimos cuadrados o máxima verosimilitud. En el caso de distribución normal de errores ambos métodos coinciden.

Denominando $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$ al valor que el modelo estimado predice para la observación y_i, el error cometido en esa previsión es:

$$e_i = y_i - \widehat{y_i} = y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)$$

Es importante notar que $\widehat{\beta_0}$ y $\widehat{\beta_1}$ son los valores estimados del modelo. En el Ejemplo 1 estos valores son $-100,22$ y 0.97 respectivamente. Nuestro objetivo es calcular $\widehat{\beta_0}$ y $\widehat{\beta_1}$. El criterio de mínimos cuadados asigna a $\widehat{\beta_0}$ y $\widehat{\beta_1}$ el valor que minimiza la suma de errores al cuadrado de todas las observaciones.

$$S = \min(\sum_{i=1}^{n} e_i^2) = \min(\sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x)^2$$

Para calcular el mínimo de esta ecuación, hay que derivar respecto de $\widehat{\beta_0}$ y $\widehat{\beta_1}$. La solución que se obtiene es:

$$\widehat{\beta_1} = \frac{cov(x,y)}{var(x)} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Utilizando las fórmulas anteriores se obtienen los valores de $\widehat{\beta_0}$ y $\widehat{\beta_1}$. Para estimar la varianza se utiliza la varianza residual:

$$\widehat{s}_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

2.3 Intervalos y contrastes

Una vez estimados los valores de los parámetros, es necesario obtener una medida de la precisión con que se han estimado estos coeficientes. Este problema lo resolveremos mediante la construcción de *Intervalos de confianza* para los coeficientes.

También será necesario contrastar si es posible que el valor auténtico del parámetro sea alguno determinado. Este problema lo resolveremos mediante el *Contraste de hipótesis*.

2.3.1 Intervalos de confianza.

Un intervalo de confianza proporciona un conjunto de valores que, con un nivel de confianza predeterminado, contendrá el auténtico valor del parámetro. Tanto para el cálculo de intervalos de confianza, como para el de contrastres de hipótesis es preciso estimar el error estándar en la estimación del parámetro que se quiere contrastar.

Llamaremos $SE(\widehat{\beta}_i)$ al error estándar en la estimación de $\widehat{\beta}_i$. Puede demostrarse que

$$SE(\widehat{\beta}_i) = \sqrt{\frac{\widehat{s}_R^2}{ns_x^2}}$$

Esta medida, proporcionada por cualquier paquete informático, permite construir intervalos de confianza y contrastes de hipótesis. Así, el intervalo de confianza para un nivel de confianza $1-\alpha$ para $\widehat{\beta}_1$ será:

$$\beta_1 \in \widehat{\beta}_1 \pm t_{\alpha/2} SE(\widehat{\beta}_1)$$

donde $t_{\alpha/2}$ es el valor de la distribución t con $n-2$ grados de libertad que deja a su derecha $\alpha/2$ de probabilidad. Si se tienen más de 25 o 30 observaciones y se quiere un nivel de confianza del 95%, puede aproximarse el intervalo anterior mediante:

$$\beta_1 \in \widehat{\beta}_1 \pm 2SE(\widehat{\beta}_1)$$

Esta expresión permite razonar que el valor auténtico de β_1 se encuentra con confianza 95% entre los valores $(\widehat{\beta}_1 - 2SE(\widehat{\beta}_1), \widehat{\beta}_1 + 2SE(\widehat{\beta}_1))$. La información proporcionada por el intervalo de

confianza es muy útil. Si el intervalo es muy ancho, la precisión de la estimación es baja. Por el contrario, un intervalo estrecho indica que tenemos una estimación muy precisa.

El intervalo permite también realizar contrastes de hipótesis. Así, por ejemplo si queremos contrastar la hipótesis de que el auténtico valor de β_1 es cero, podemos ver si el intervalo de confianza contiene o no el valor cero. En caso de que el valor 0 esté contenido podemos concluir que es posible que $\beta_1 = 0$. En caso de que el valor cero no esté comprendido en el intervalo, concluiremos que el cero es un valor incompatible con nuestras observaciones.

El intervalo de confianza permite por tanto obtener una serie de valores de β_1 que son compatibles con nuestras observaciones.

2.3.2 Contraste de hipótesis.

Una forma alternativa de resolver el problema anterior es mediante el contraste de hipótesis.

El contraste de hipótesis nos permite contestar a preguntas como:

1. ¿ Es un valor admisible de β_1 el valor β_1^*?
2. ¿ Es realmente significativa (influye) la variable x sobre la variable y?
3. Se sabe de investigaciones anteriores que $\beta_1$vale β_1^*. ¿He obtenido un resultado similar teniendo en cuenta que a mí me sale $\widehat{\beta}_1$?

El planteamiento es similar en cualquiera de los tres casos. Se plantea la *Hipótesis Nula*, que denominaremos H_0, y la *Hipótesis Alternativa*, H_1. A continuación, vemos cual de las dos hipótesis es más cercana a nuestros datos.

El contraste t resuelve los problemas planteados en el inicio de la sección. El planteamiento es muy simple. Se puede demostrar que

$$\frac{\widehat{\beta}_1 - \beta_1}{SE(\widehat{\beta}_1)} = t$$

se distribuye como una t de $n - 2$ grados de libertad. Entonces, sabiendo que $\widehat{\beta}_1$ es el valor que hemos obtenido de la estimación y que $SE(\widehat{\beta}_1)$ es el error estándar estimado, no hay más que sustituir el valor β_1 auténtico por el que queremos comprobar. El valor t obtenido deberá ser una t de $n - 2$ grados de libertad. Como la distribución t es conocida, se puede inferir si el valor obtenido proviene verosímilmente de una distribución t o no.

Ejemplo 2:

Los datos que se van a estudiar son el número de varones y mujeres en 104 municipios sevillanos obtenidos del censo de Floridablanca. La figura muestra el número de varones y mujeres en los municipios de Sevilla. Como puede observarse, el número de varones y de mujeres está relacionado, y esta relación es de tipo lineal, por lo que es razonable ajustar una recta de regresión.

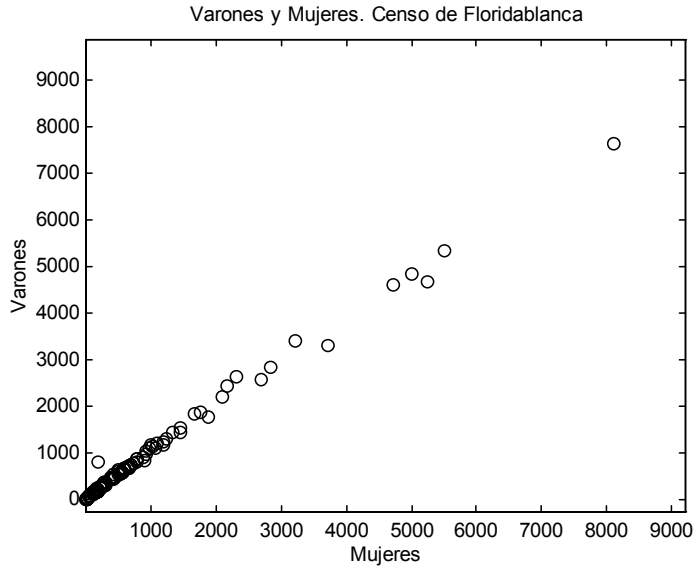


Figura 3: Número de Hombres y mujeres en municipios de Sevilla. Censo de Floridablanca.

A continuación se muestra el resultados del análisis de regresión simple para estos datos. La Ecuación de regresión obtenida es:

$$\text{Varones} = \underset{(31.4)}{77.79} + \underset{(0.007)}{0.93}\text{Mujeres}$$

Los números entre paréntesis son los errores estándar estimados $SE(\widehat{\beta}_i)$.

Vamos a contrastar dos hipótesis. En primer lugar vamos a contrastar si el número de mujeres en un pueblo proporciona información sobre el número de hombres. Esto equivale a contrastar si el valor auténtico de β_1 es igual a cero. El planteamiento es:

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

Construimos el estadístico *t*,

$$\frac{0.93 - 0}{0.007} = 137.54$$

El valor obtenido es muy grande para una *t* de 102 grados de libertad, que proporciona valores entre -2 y +2 con un 95% de probabilidad. Concluimos por tanto que β_1 no puede ser igual a cero. Es interesante observar que el contraste de falta de significatividad del coeficiente equivale a dividir el coeficiente estimado por su error estándar. Ese cociente, que en nuestro ejemplo es 137.54, recibe el nombre de estadístico *t*, y los paquetes informáticos lo porporcionan automáticamente. Es estadístico *t* sirve para comprobar rápidamente si la variable *x* es significativa ya que si su valor está comprendido entre -2 y +2 pesaremos que no lo es, y en caso contrario concluiremos que *x* si influye.

Otro contraste interesante en nuestro ejemplo es comprobar si un incremento de 100 mujeres conlleva un incremento de 100 hombres. Esto equivale a contrastar si β_1 puede valer 1. El planteamiento es análogo al del caso anterior:β

$$H_0 : \ \beta_1 = 1$$
$$H_1 : \ \beta_1 \neq 1$$

Construimos el estadístico *t*,

$$\frac{0.93 - 1}{0.007} = -10$$

De nuevo, el valor obtenido es muy grande para una *t* de 102 grados de libertad, por lo que concluimos que, aunque aparentemente el coeficiente .93 está muy cerca de 1, en realidad es distinto de 1. Obsérvese que este resultado es coherente con la realidad, ya que existen menos varones que mujeres en la población, debido a la mayor esperanza de vida de las mujeres.

Si hubiéramos construido un intervalo de confianza:

$$\beta_1 \in 0.93 \pm 2 \bullet 0.007$$

es decir que β_1 estará comprendida en el intervalo 0.9160 0.9314 con una confianza del 95%. Indudamente el intervalo de confianza no contiene el valor 0 ni el valor 1 que hemos rechazado como posibles en los contrastes de hipótesis.

3. Contraste de regresión.

El contraste de regresión es una forma alternativa de resolver el contraste de significatividad. Su uso en regresión simple no aporta nada nuevo al contraste *t*. Sin embargo en regresión múltiple va a ser un arma muy valiosa.

Se puede demostrar que si $\beta_1 = 0$, el cociente

$$F = \frac{VE}{VNE/n-2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2} = \frac{n\hat{\beta}_1^2 s_x^2}{\hat{s}_R^2}$$

se distribuye como una $F_{1,n-2}$. El uso del estadístico *F* es inmediato. En el ejemplo anterior, *F* vale 18900, que indica que la hipótesis nula puede rechazarse.

4. Coeficiente de determinación.

El coeficiente de determinación, R^2, proporciona la cantidad de variabilidad de *y* que explica la *x*. Se define

$$R^2 = \frac{VE}{VT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{ns_y^2}$$

El coeficiente de determinación proporciona información sobre si *x* e *y* están muy relacionadas o no. En el ejemplo del censo de Floridablanca, el valor de $R^2 = 99.46$. Esto indica que el número de hombres y de mujeres está muy relacionado.

5. Predicción.

Una aplicación inmediata del análisis de regresión, es utilizar la recta estimada para predecir valores de y. Se distinguen dos tipos de predicciones:

1. Predicción del valor promedio de y para un valor determinado de x.
2. Predicción de una observación concreta.

La predicción del valor promedio de y cuando x toma determinado valor se obtiene substituyendo x en la ecuación de regresión estimada. Así, en el ejemplo del Censo de Floridablanca, si queremos saber cuantos hombres tendrán en promedio los municipios con 2000 mujeres, hombres$=77,79 + .93 \bullet 2000 = 1937.8$.

Este valor, hay que completarlo con un intervalo de confianza que indicará para un nivel de confianza determinado (Normalmente el 95%) en que región se puede encontrar el número promedio de hombres en los pueblos que tienen 2000 mujeres.

En general, se puede demostrar que cuando $x = x_0$, el promedio de y, es decir $y_0 = \widehat{\beta_0} + \widehat{\beta_1} x_0$, se encuentra con confianza 95% en el intervalo

$$y_0 \pm 2\widehat{s}_R \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (\widehat{x}_i - \bar{x})^2}}$$

La predicción de una observación concreta es un problema muy similar. En este caso se pretende saber cuántos hombres habrá en un pueblo concreto con 2000 mujeres. El valor previsto es, como en el caso anterior hombres$=77,79 + .93 \bullet 2000 = 1937.8$. Sin embargo el intervalo, que aquí llamaremos intervalo de predicción, varía y toma el valor:

$$y_0 \pm 2\widehat{s}_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (\widehat{x}_i - \bar{x})^2}}$$

Obsérvese que como es lógico el intervalo de predicción es mayor que el intervalo de confianza, ya que hay más incertidumbre en la predicción de una observación que en la predicción de un valor promedio.

6. Diagnosis en regresión simple.

Una vez ajustado el modelo es necesario comprobar que se cumplen las hipótesis que hemos realizado. Será necesario comprobar la linealidad de los datos, la homocedasticidad, y la normalidad.

La comprobación de las propiedades anteriores se realiza mediante gráficos. El más importante es el gráfico de residuos frente a valores ajustados. La Figura 4 muestra el gráfico de Residuos frente a valores ajustados para el ejemplo de pesos y alturas.

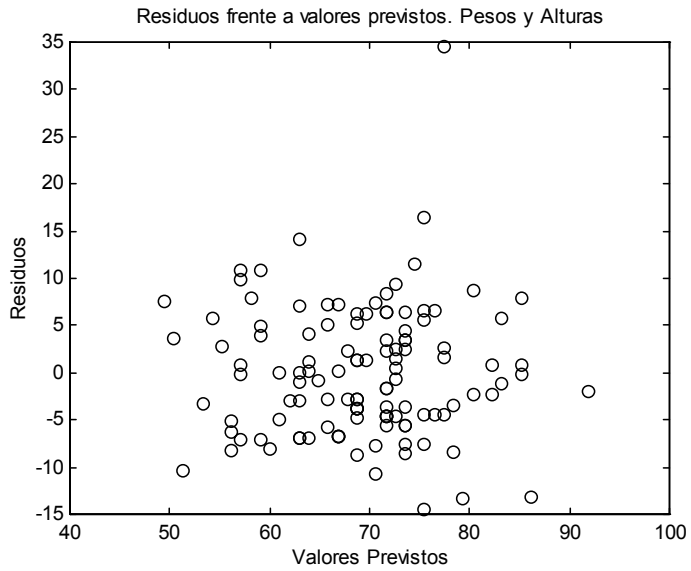


Figura 4: Residuos de la regresión de Pesos y Alturas de la Figura 2.

En la figura puede observarse que los residuos no presentan ninguna estructura especial. Este gráfico debe presentar un aspecto totalmente aleatorio, sin ninguna estructura.

Cuando los datos no son lineales, el gráfico X-Y suele presentar una estructura como la de la figura 5. Estos datos son simulados, y en ellos se aprecia la estructura no lineal.

Si a estos datos se les ajusta una recta de regresión, la recta deja muchos puntos por encima en la zona de *x* baja y alta, mientras para *x* intermedias los puntos quedan en general por debajo de la recta de regresión. Cuando se mira el gráfico de los resíduos, se aprecia con mucha más claridad la falta de linealidad de los datos.

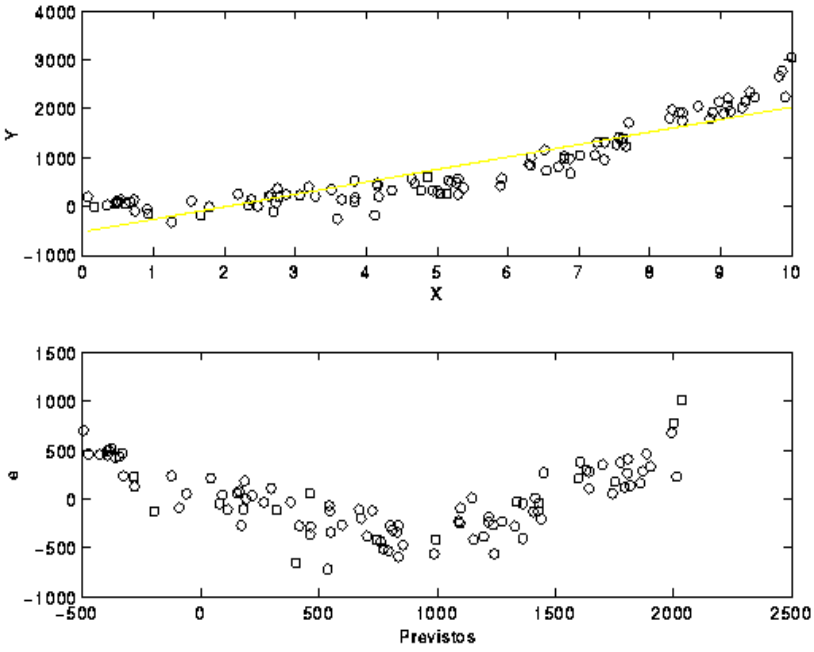


Figura 5: Datos no lineales. Recta de regresión y residuos.

Al analizar los resíduos, es necesario observar también si existe heterocedasticidad. La figura 6 presenta un ejemplo de datos X-Y heterocedásticos. Como puede comprobarse, la variablidad de los datos aumenta al incrementarse el valor de X. Si, tras ajustar la recta de regresión, observamos el gráfico de Residuos frente a valores ajustados, vemos cómo es patente la heterocedasticidad de los datos.

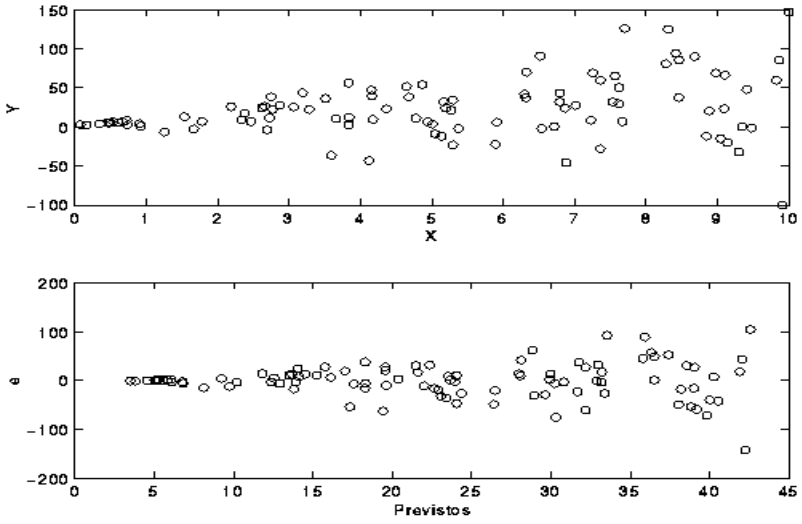


Figura 6: Datos heterocedásticos y sus resíduos.

La hipótesis de independencia es muy importante. Si no se cumple, el análisis de regresión es erróneo y puede llevar a resultados equívocos. Existen contrastes para determinar si los datos son independientes, como el contraste de Durbin-Watson. En cualquier caso, si los datos no son independientes, como ocurre con las series temporales, es mejor utilizar otro tipo de técnicas.

Finalmente es interesante realizar un histograma de los residuos para comprobar que presentan un aspecto razonablemente Normal.

7. Transformaciones.

Cuando las hipótesis del modelo no se cumplen es necesario transformar los datos, de manera que los datos transformados cumplan las hipótesis. Las transformaciones más utilizadas son:

1. Logaritmo: $y' = \log y$ o $x' = \log x$. La transformación logrítmica puede aplicarse a las variables x e y. Esta transformación elimina en ocasiones los problemas de falta de linealidad

o heterocedasticidad.
Cuadrado: $y' = y^2$ o $x' = x^2$.

2. Inversa: $y' = \frac{1}{y}$ o $x' = \frac{1}{x}$.
3. Raíz cuadrada: $y' = \sqrt{y}$ o $x' = \sqrt{x}$. Esta transformación es muy útil cuando los datos proceden de una distribución de Poisson, es decir representan número de sucesos .

Regresión Múltiple

8. Introducción.

La regresión múltiple es una extensión inmediata de la regresión simple. En aquella se pretende estudiar la relación entre una variable dependiente y, y una variable independiente x. En la práctica es claro que y puede depender de más de una variable independiente. Por ejemplo en el caso de Pesos y Alturas, es posibe que el peso dependa de la altura y de determinadas condiciones genéticas. La extensión es por tanto inmediata. El modelo que se plantea en regresión múltiple será por tanto

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i \qquad (3)$$

donde $x_1, x_2, \ldots, x_k$ son las variables independientes o explicativas. Las variables explicativas pueden ser cuantitativas como es el caso de la Altura, o cualitativas como por ejemplo el sexo de la persona.

9. El Modelo.

El modelo de regresión múltiple requiere, como en el caso de regresión simple, hacer las siguientes hipótesis:

1. **Linealidad**: Es una hipótesis fundamental. Para ajustar un modelo como el de la ecuación (3) a un conjunto de datos es preciso que éstos cumplan esa ecuación. En el caso de regresión simple, los datos debían tener un aspecto razonablemente recto. Si hay dos variables explicativas, la ecuación (3) queda reducida a

   $$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

   que es la ecuación de un plano. Los datos deben por tanto ser planos. Si existen más de dos variable explicativas, la ecuación (3) es un hiperplano y no podemos visualizar el aspecto de los datos.
2. $\mathbf{E}(\mathbf{u}_i) = 0$: El valor promedio del error es cero. Esta hipótesis implica que el ajuste se va realizar está centrado respecto a los datos. Así, en virtud de esta hipótesis cabe esperar que el plano o hiperplano de regresión esté centrado en la nube de puntos de los datos.
3. $var(u_i) = \sigma^2 =$**cte**: La varianza de los errores es constante y no depende del nivel de las variables. Esta hipótesis implica que la nube de puntos de los datos tiene siempre una anchura semejante. Si lo datos cumplen esta hipótesis se dice que son **Homocedástico**s. Por el contrario, datos cuya variablidad no es constante se denominana **Heterocedásticos**.
4. **Independencia** $\mathbf{E}(\mathbf{u}_i\mathbf{u}_j) = 0$ Esta hipótesis implica que las observaciones son independientes. Es decir una observación no ofrece información sobre los valores de la siguiente.
5. **Normalidad** $u_i \sim N(0, \sigma^2)$ Esta hipótesis se refiere a que los errores se distribuyen normalmente, es decir siguiendo una campana de Gauss.

10. Ajuste.

El ajuste del modelo se realiza por mínimos cuadrados o máxima verosimilitud. En el caso de distribución normal de errores ambos métodos coinciden.

Denominando $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1}x_{1i} + \widehat{\beta_2}x_{2i} + \cdots + \widehat{\beta_k}x_{ki}$ al valor que el modelo estimado predice para la observación y_i, el error cometido en esa previsión es:

$$e_i = y_i - \widehat{y_i} = y_i - (\widehat{\beta_0} + \widehat{\beta_1}x_{1i} + \widehat{\beta_2}x_{2i} + \cdots + \widehat{\beta_k}x_{ki})$$

donde $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_k}$ son los valores estimados del modelo. El criterio de mínimos cuadados asigna a $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_k}$ el valor que minimiza la suma de errores al cuadrado de todas las observaciones.

$$S = \min(\sum_{i=1}^{n} e_i^2) = \min(\sum_{i=1}^{n}(y_i - (\widehat{\beta_0} + \widehat{\beta_1}x_{1i} + \widehat{\beta_2}x_{2i} + \cdots + \widehat{\beta_k}x_{ki}))^2$$

Para calcular el mínimo de esta ecuación, hay que derivar respecto de $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_k}$. La solución que se obtiene debe expresarse en términos matriciales y es:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad (4)$$

donde $\widehat{\boldsymbol{\beta}}$ representa un vector columna de dimensión $k + 1$ que contiene los parámetros

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta_1} \\ \widehat{\beta_2} \\ \vdots \\ \widehat{\beta_k} \end{pmatrix}$$

$\mathbf{X}$ es la denominada matriz de diseño, de dimensión $n \times (k + 1)$.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}$$

La varianza residual tiene la expresión

$$\widehat{s}_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - k - 1}$$

11. Intervalos y contrastes

Tal como hacíamos en regresión simple, una vez estimados los valores de los parámetros, es necesario obtener una medida de la precisión con que se han estimado estos coeficientes. Este problema lo resolveremos mediante la construcción de *Intervalos de confianza* y *Contrastes de hipótesis*.

11.1 Intervalos de confianza.

Un intervalo de confianza proporciona un conjunto de valores que, con un nivel de confianza predeterminado, contendrá el auténtico valor del parámetro. Tanto para el cálculo de intervalos de confianza, como para el de contrastres de hipótesis es preciso estimar el error estándar en la estimación del parámetro que se quiere contrastar. La estimación de los errores estándar en regresión múltiple tiene una expresión compleja. Llamando $\widehat{\sum}$ a la matriz definida por

$$\widehat{\sum} = (\mathbf{X}'\mathbf{X})^{-1}\hat{s}_R^2 = \begin{pmatrix} q_{11} & \cdots & q_{k+1,1} \\ \vdots & \ddots & \vdots \\ q_{1k+1} & \cdots & q_{k+1,k+1} \end{pmatrix}$$

puede demostrarse que $SE(\widehat{\beta}_i)$, error estándar en la estimación de $\widehat{\beta}_i$, es igual a la raíz cuadrada del elemento de la diagonal principal de la matriz $\widehat{\sum}$. En cualquier caso, los valores de los errores estándar son proporcionados por cualquier paquete de software estadístico.

Con esta medida, podemos construir intervalos de confianza y contrastes de hipótesis. Así, el intervalo de confianza para un nivel de confianza $1 - \alpha$ para $\widehat{\beta}_i$ será:

$$\beta_i \in \widehat{\beta}_i \pm t_{\alpha/2} SE(\widehat{\beta}_i)$$

donde $t_{\alpha/2}$ es el valor de la distribución t con $n - k - 1$ grados de libertad que deja a su derecha $\alpha/2$ de probabilidad. Si se tienen más de 25 o 30 observaciones y se quiere un nivel de confianza del 95%, puede aproximarse el intervalo anterior mediante:

$$\beta_i \in \widehat{\beta}_i \pm 2SE(\widehat{\beta}_i)$$

Esta expresión permite razonar que el valor auténtico de β_i se encuentra con confianza 95% entre los valores $(\widehat{\beta}_i - 2SE(\widehat{\beta}_i), \widehat{\beta}_i + 2SE(\widehat{\beta}_i))$. La información proporcionada por el intervalo de confianza es muy útil. Si el intervalo es muy ancho, la precisión de la estimación es baja. Por el contrario, un intervalo estrecho indica que tenemos una estimación muy precisa.

El intervalo permite también realizar contrastes de hipótesis. Así, por ejemplo si queremos contrastar la hipótesis de que el auténtico valor de β_i es cero, podemos ver si el intervalo de confianza contiene o no el valor cero. En caso de que el valor 0 esté contenido podemos concluir que es posible que $\beta_i = 0$. En caso de que el valor cero no esté comprendido en el intervalo, concluiremos que el cero es un valor incompatible con nuestras observaciones.

El intervalo de confianza permite por tanto obtener una serie de valores de β_i que son compatibles con nuestras observaciones.

11.2 Contraste de hipótesis sobre coeficientes individuales.

Una forma alternativa de resolver el problema anterior es mediante el contraste de hipótesis.

El planteamiento es similar al realizado en regresión simple. Se plantea la *Hipótesis Nula*, que denominaremos H_0, y la *Hipótesis Alternativa*, H_1. A continuación, vemos cual de las dos hipótesis es más cercana a nuestros datos.

De nuevo el contraste *t* resuelve los problemas planteados en el inicio de la sección. El planteamiento es muy simple. Se puede demostrar que

$$\frac{\widehat{\beta}_i - \beta_i}{SE(\widehat{\beta}_i)} = t$$

se distribuye como una *t* de $n - k - 1$ grados de libertad. Entonces, sabiendo que $\widehat{\beta}_i$ es el valor que hemos obtenido de la estimación y que $SE(\widehat{\beta}_i)$ es el error estándar estimado, no hay más que sustituir el valor β_i auténtico por el que queremos comprobar. El valor *t* obtenido deberá ser una *t* de $n - k - 1$ grados de libertad. Como la distribución *t* es conocida, se puede inferir si el valor obtenido proviene verosímilmente de una distribución *t* o no.

11.3 Contraste de regresión F.

El contraste de regresión en regresión múltiple sirve para comprobar si el modelo explica una parte significativa de la variabilidad de *y*.

Se puede demostrar que si $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, el cociente

$$\frac{VE/k-1}{VNE/n-k-1} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2} = F$$

se distribuye como una $F_{k-1,n-k-1}$.Si el valor obtenido es un valor probable para una $F_{k-1,n-k-1}$ llegaremos a la conclusión de que el modelo no explica conjuntamente nada. Si por el contrario el test indica que el valor obtenido no puede razonablemente provenir de una *F*, entonces el modelo explica una parte significativa de *y*.

12. Multicolinealidad.

En ocasiones el analista pretende explicar la variable *y* mediante una serie de variables $x_1, x_2, \ldots, x_k$ que tienen mucha relación entre sí. Si ésto ocurre, el modelo no va a poder distinguir qué parte de *y* es explicada por una variable y qué parte de *y* es explicada por otra. Vamos a ilustrar el caso con un ejemplo.

Ejemplo:

Los datos que se presentan en la figura 7 son el número de accidentes de tráfico en una serie de provincias españolas (*Acci*), el número de permisos de conducir (*Perm*) y el número de vehículo matriculados en las mismas provincias (*Matric*).

La figura7 muestra la relación entre los accidentes y permisos y matrículas. Es evidente que la relación existe y además es muy fuerte. A medida que aumenta el número de permisos o de matrículas el número de accidentes también lo hace.

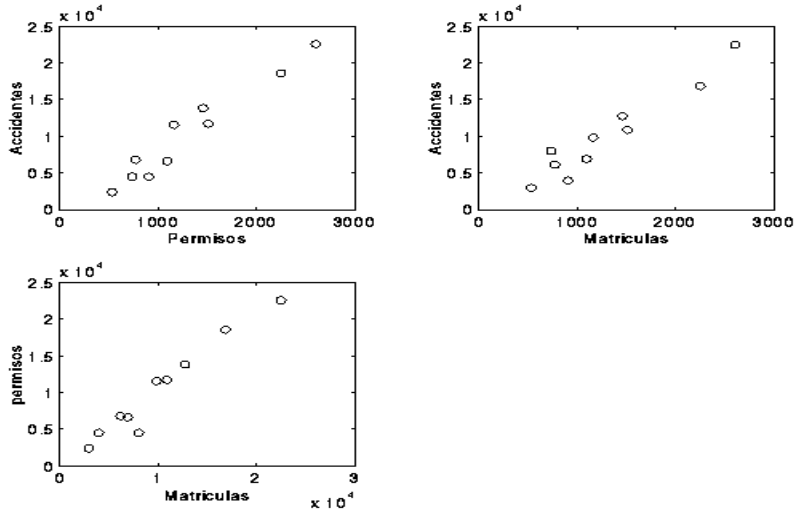


Figura 7: Datos de accidentes de automóvil en provincias españolas.

La regresión obtenida es

$$\begin{array}{lll} y = & 250.63 & +0.03Perm \quad +.007Matric \\ SE & & (0.04) \qquad (0.04) \\ t & & .69 \qquad 1.83 \end{array}$$

donde, aparentemente ninguna de las dos variables influye sobre el número de accidentes, ya que el estadístico t es muy bajo. Esta paradoja, por la que variables aparentemente muy significativas no resultan significativas en una regresión múltiple, es a veces debida a la **multicolinealidad**.

Las regresiones simples entre y y las variables x son:

$$\begin{array}{lll} y = & 278.24 & +0.1Matric \\ SE & & (0.0085) \\ t & & 11.68 \end{array} \qquad (5)$$

$$\begin{array}{lll} y = & 216.48 & +0.1Perm \\ SE & & (0.01) \\ t & & 9.81 \end{array} \qquad (6)$$

$$\begin{array}{lll} Matric = & -470.7 & +Perm \\ SE & & (0.09) \\ t & & 12.4 \end{array} \qquad (7)$$

que indican que sobre y influyen tanto *Perm* como *Matric* como era de esperar a la vista de la figura.(Ecuaciones 5 y 6). Esta aparente paradoja es un ejemplo claro de multicolinealidad. Como muestra la ecuación (7), el número de vehículos matriculados depende a su vez mucho del número de permisos que hay en cada provincia, y el modelo no puede discernir qué parte de la variabilidad de y es debida a cada una de las dos variables. Este fenómeno es muy frecuente, pues es natural tener variables alternativas o muy correladas entre ellas.

La multicolinealidad se detecta porque **aunque el contraste t dé un valor muy bajo**, **el contraste conjunto F indica que una parte importante de la variabilidad es explicada por el modelo**. En nuestro ejemplo, el contraste F vale 64.06 que es una valor muy improbable para una $F_{2,7}$. Un forma adicional de comprobar la existencia de multicolinealidad es que el coeficiente de determinación es alto aunque las variables no sean significativas.

13. Coeficiente de determinación.

El coeficiente de determinación, R^2, proporciona la cantidad de variabilidad de y que explica la x. Se define al igual que en regresión simple

$$R^2 = \frac{VE}{VT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{ns_y^2} \qquad (8)$$

El coeficiente de determinación proporciona información sobre si x e y están muy relacionadas o no. Sin embargo, el coeficiente de determinación definido en 8 tiene el problema de que al incluir nuevas variable aumenta su valor, *incluso cuando esas variables no sean significativas*. Este problema hace que R^2 no se pueda utilizar como criterio válido para incluir o excluir variables.

Para evitar este problema se define el Coeficiente de Determinación corregido por grados de libertad, $\overline{R}^2$. Se define $\overline{R}^2$ como

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k+1}$$

Este coeficiente $\overline{R}^2$ no tiene los inconvenientes de R^2 ya que al introducir más variables en el modelo no aumenta necesariamente su valor.

13. Diagnosis.

La diagnosis en regresión múltiple es más compleja que en regresión simple. Esto es debido a que no podemos visualizar los datos correctamente.

Para comprobar las hipótesis de linealidad, heterocedasticidad y normalidad se realiza como en regresión simple gráficos de residuos frente a valores ajustados.

Además, en regresión múltiple es útil realizar gráficos de resíduos frente a las variables explicativas x. En estos gráficos es posible identificar si alguna variable produce los efectos de falta de linealidad y hetercedasticidad.

Ejemplo

Se va a construir un modelo para explicar cómo influye sobre las millas recorridas por un automóvil con un galón de gasolina (MpG) la aceleración del coche (Aceleración), la potencia (Potencia) y el precio (Precio).

La figura 8 muestra que la relación entre MpG y Potencia es fuerte pero no es lineal. Entre las denás variables la relación es mucho más suave. Esta falta de linealidad nos indica que muy probablemente será necesario transformar los datos.

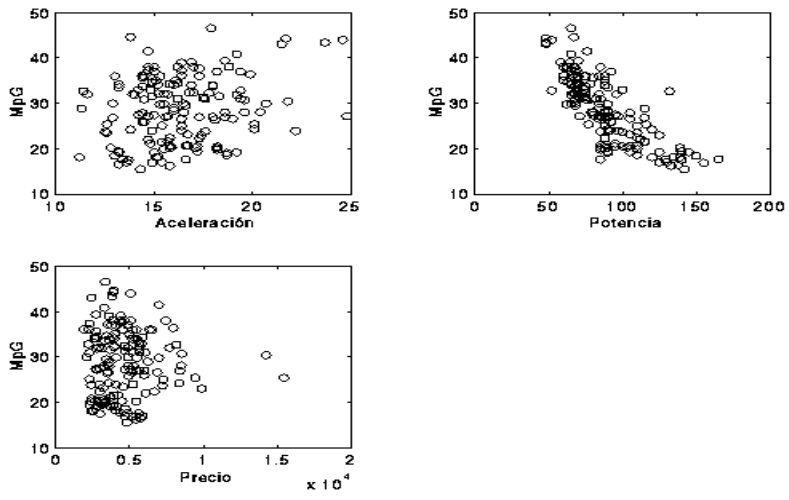


Figura 8: Relación entre Consumo, Potencia y Precio de automóviles

Ajustando un modelo a los datos sin transformar, obtenemos:

MpG	=61.69	-0.27*Potencia*	–0.64*Aceleracion*	+0.000357Precio
SE		(0.02)	(0.16)	(0.0002)
t		-16.22	-3.95	2.02

$\overline{R}^2 = 0.656$

Los estadísticos *t* indican que las tres variables son significativas, ya que son mayores que 2 en valor absoluto. Los signos de las variables indican:

- Un incremento de Potencia de 1 HP produce un descenso de 0,27 millas recorridas por galón. Para que esto sea cierto tanto la aceleración como el precio deben permanecer constantes.
- Un incremento de aceleración de uns unidad produce un un descenso de 0,67 millas recorridas por galón. Para que esto sea cierto tanto la potencia como el precio deben permanecer constantes.
- Un incremento de precio de 1 unidad, produce un incremento de la longitud recorrida por galón de 0.00035. Siempre que potencia y aceleración sean constantes.

Este resulatado es lógico, pues los coches más caros, con mecánica más depurada consumen menos a igual potencia y aceleración.

El gráfico de residuos frente a valores ajustados se presenta en la figura. Como puede observarse y tal como era previsible teniendo en cuenta la falta de linealidad de los datos, los residuos presentan estructura. Son heterocedásticos y posiblemete no lineales.

Para tratar de arreglar este problema se van a transformar los datos a logaritmos.

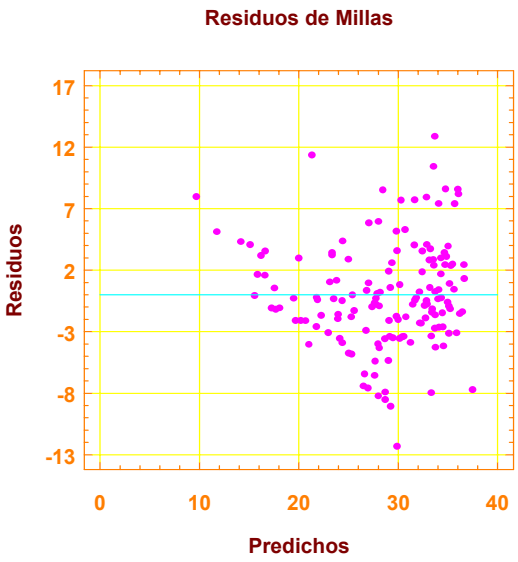


Figura 9: Residuos del modelo. Datos

La figura muestra los gráficos de dispersión para los datos en logaritmos. La mejora de la linealidad es evidente. A pesar de todo, la relación entre el consumo y el precio no es demasiado clara. Esto no quiere decir que no exista, sino que al actuar tanto sobre el precio, como sobre el consumo otros factores, la relación puede quedar oculta. La regresión múltiple nos dirá si existe o no la relación.

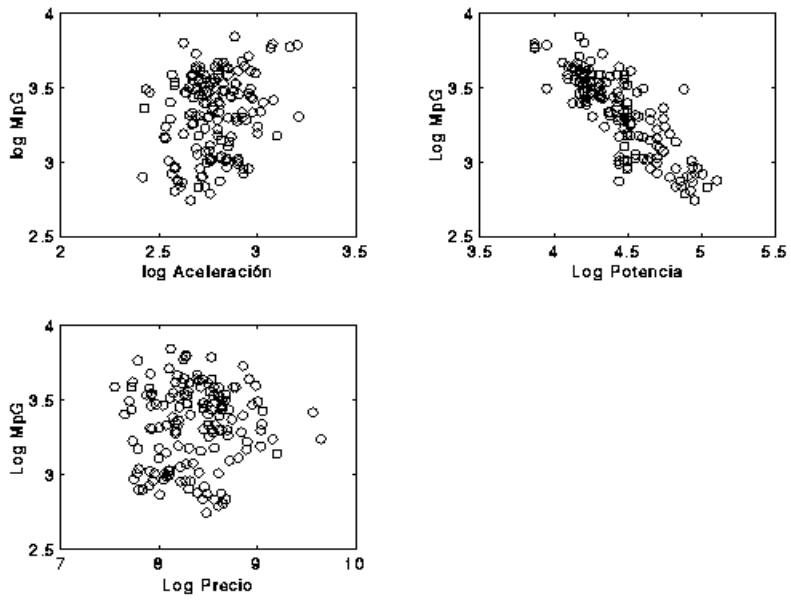


Figura 10: Relaciones en logaritmos. Obsérvese la linealidad entre log(MpG) y

Log(MpG)	=8.23	-1.01Log(*Potencia*)	–0.53Log(*Aceleracion*)	+0.12Log(Precio)
SE		(0.05)	(0.08)	(0.03)
t		-20.6	-6.2	4.37

$\overline{R}^2 = 0.7529$

Como puede apreciarse, hemos incrementado la variabilidad explicada de forma notable. Los efectos en logaritmos tienen un significado muy concreto, ya que representan incrementos porcentuales de y cuando x se incrementa porcentualmente. En efecto, los efectos estimados son:

- Si la potencia se incrementa en un 10%, las millas recorridas se reducen en un $10 \bullet 1.01 = 10.1\%$. Manteniendo constantes las demás variables.
- Si la aceleración aumenta un 10%, las millas recorridas se reducen en un $10 \bullet 0.53 = 5.3\%$. Manteniendo constantes las demás variables
- Si el precio aumenta un 10%, las millas recorridasaumentan en un $10 \bullet 0.12 = 1.2\%$. Manteniendo constantes las demás variables

La figura muestra el gráfico de residuos frente a valores ajustados de este modelo. En ella vemos que los residuos no presentan estructura, por lo que el modelo logarítmico va a resultar adecuado.

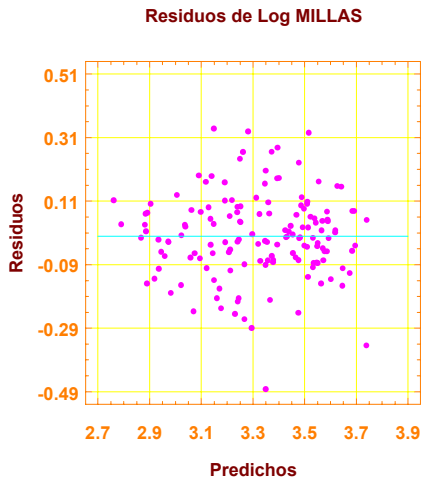


Figura 11: Residuos del modelo en logaritmos.

13.1 Gráficos especiales de residuos

En regresión múltiple la visualización de los datos es muy compleja, ya que se encuentran en un espacio de k+1 dimensiones.

Como método alternativo de visualización de los datos se puede realizar un gráfico de variable añadida. Este gráfico pretende mostrar la relación entre y y una variable x, limpia de los efectos de las demás variables. Supongamos que se desea analizar un modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3_i} + u_i$$

y deseamos visualizar la relación entre y y, por ejemplo, x_3.

La construcción es como sigue:

1. Realizamos la regresión entre y y el resto de las variables x.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

y obtenemos los residuos de la regresión: $e_i = y_i - \hat{y}_i$. Estos residuos mantienen la relación entre y y x_3. Estos residuos se denominan $e_{y.x1,x2.}$

2. Realizamos la regresión entre x_3 y las demás variables x.

$$\hat{x}_{3i} = \widehat{\beta *}_0 + \widehat{\beta *}_1 x_{1i} + \widehat{\beta *}_2 x_{2_i}$$

Los residuos de esta expresión $e *_i = x_{3i} - \widehat{x_{3\,i}}$ se denominan $e_{x3.x1,x2.}$, y representan la variable x_3 limpia de los efectos de x_1 y x_2.

Veamos un ejemplo.

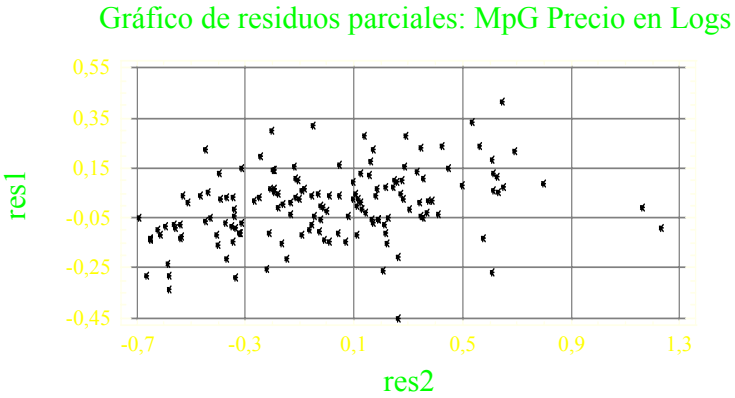
Ejemplo:

Se va realizar el gráfico de variable añadida entre Log MpG y Log Precio con los datos del ejemplo anterior.

Para ello calculamos en primer lugar la regresión entre Log Mpg como variable dependiente, y Log Aceleración y Log Potencia como independientes. Los residuos de esta ecuación los denominaremos RES1.

A continuación realizamos la regresión entre Log Precio como variable independiente, y Log Aceleración y Log Potencia como independientes. Los residuos de esta ecuación los denominaremos RES2.

El gráfico de dispersión entre RES1 y RES2 es gráfico de residuos que se presenta a continuación:



En el gráfico se observa la relación positiva que se obtiene en la regresión múltiple. Se puede demostrar que si se ajustase una recta de regresión a los datos de la figura se obtendría la misma $\widehat{\beta}$ que se obtiene en la regresión múltiple.

14. Variables cualitativas.

En ocasiones el análisis de regresión requiere diferenciar la relación entre las variables x e y según algúna variable cualitativa. Así, por ejemplo, la figura 12 muestra la relación entre dos variables x e y. Es evidente que existen dos grupos de observaciones. Si no tuviéramos en cuenta

la existencia de los dos grupos de observaciones, el ajuste del modelo de regresión va a ser erróneo.

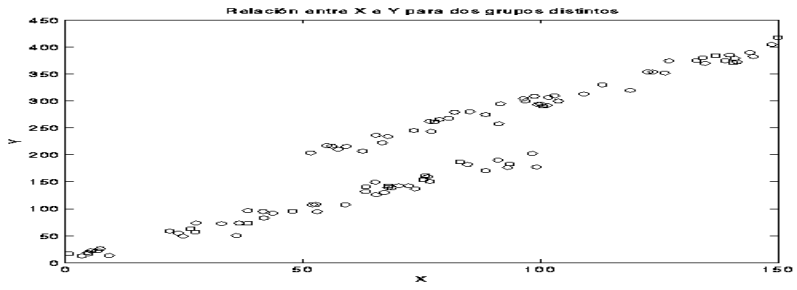


Figura 12: Relación entre las variable *x* e *y* cuando existen dos grupos.

En efecto, supongamos que no consideramos el efecto del grupo. entonces, la recta de regresión que se va estimar será la que marca la figura 13.2, es decir será una recta que no representará la relación entre *x* e *y* para ninguno de los dos grupos pues va a estmarse utilizando todas las observaciones..

Si tenemos en cuenta el grupo en lugar de estimar un recta estimaremos dos, una para cada grupo. Esto se muestra en la figura 2.1

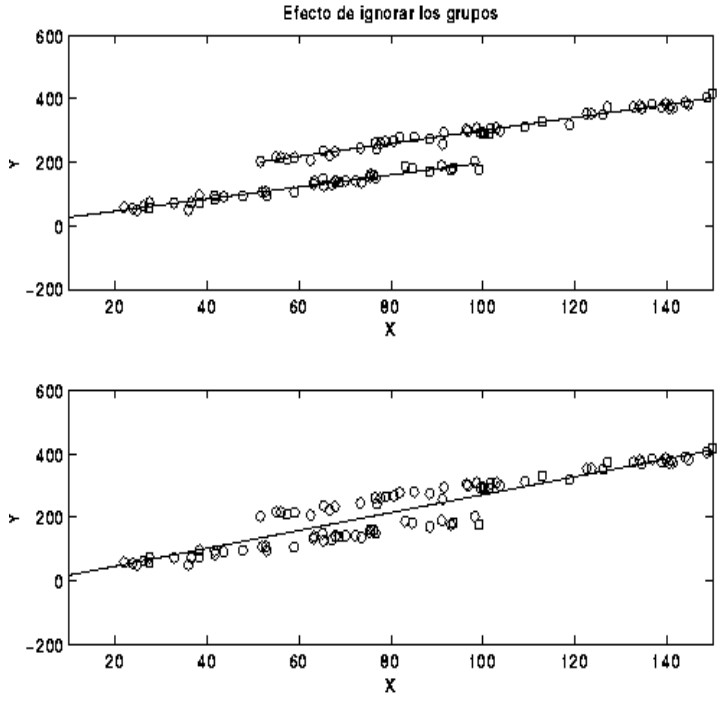


Figura 13: Rectas de regresión ajustadas teniendo en cuenta

Como se puede apreciar en las figuras, cuando existe una relación diferente entre las variables para grupos de observaciones, es preciso tener en cuenta la pertenencia a los grupos.

En la práctica este fenómeno es muy común. Ejemplos en los que se puede esperar encontrar grupos son:

- Relación entre **Pesos** y **Alturas** por **Sexos**. Cabe esperar que la relación entre peso y altura no sea la misma para hombres y mujeres.
- Relación entre **Consumo** y **Renta** para diversas familias, en función de que el cabeza de familia esté o no en **Desempleo**.

14.1 Introducción de variables dicotómicas.

Los casos que acabamos de ver se caracterizan porque los grupos son binarios, es decir que existen únicamente dos posibilidades de clasificación. Para modelizar esta situación se recurre a las variables **binarias**, **ficticias o dummies**. Se define una variable *z* que puede tomar dos valores:

$$z_i = \begin{cases} 0 \text{ si la observación } i \text{ pertenece al primer grupo} \\ 1 \text{ si la observación } i \text{ pertenece al segundo grupo} \end{cases}$$

Un vez definida la variable z, se ajusta un modelo de regresión múltiple para las variables x y z.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i$$

que tiene la propiedad de ajustar realmente dos rectas de regresión:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 x_i$$

para el grupo con $z = 1$, y

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

para el grupo con $z = 0$.

Estas dos rectas de regresión corresponden a las rectas de la figura 13.1. El modelo estima simultáneamente ambas rectas, con la particularidad, de que son paralelas.

Así por ejemplo si x es la altura, y el peso y z el sexo, podemos definir z como:

$$z_i = \begin{cases} 0 \text{ si la persona } i \text{ es hombre} \\ 1 \text{ si la persona } i \text{ es mujer} \end{cases}$$

y ajustar las rectas de regresión:

$$\hat{y}_i = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 x_i \text{ para las mujeres} \\ \hat{\beta}_0 + \hat{\beta}_1 x_i \text{ para los hombres} \end{cases}$$

Para determinar si el grupo es significativo, realizamos un contraste t sobre el coeficiente de la variable dicotómica $\hat{\beta}_2$. La realización práctica del contraste es inmediata, ya que para contrastar si β_2 es realmente igual a 0, basta con comprobar que el estadístico t del coeficiente $\hat{\beta}_2$ es mayor que 2 en valor absoluto. Evidentemente el problema puede resolverse mediante un intervalo de confianza. Si $\hat{\beta}_2 \pm 2SE(\hat{\beta}_2)$ no contiene el valor 0, pensaremos que el grupo es significativo.

Ejemplo:

Se va a ajustar una regresión a los datos de automóviles que se analizaron en regresión múltiple. Al igual que en la sección 14, se va a estudiar el consumo, Millas por Galón o MpG, de un serie de automóviles.

En el ejemplo se vió cómo el consumo dependía de la potencia en Caballos de Vapor del coche, y que la relación logarítmica era apropiada. La figura 14 muestra la relación entre el Logaritmo del consumo respecto al Logaritmo de la potencia en HP. Los puntos oscuros representan los vehículos de fabricación USA. Los puntos claros representan los vehículos no fabricados ni diseñados en Estados Unidos. Como puede observarse, existe una clara estructura en el reparto de los puntos, por lo que cabrá esperar que el origen del automóvil sea una variable significativa.

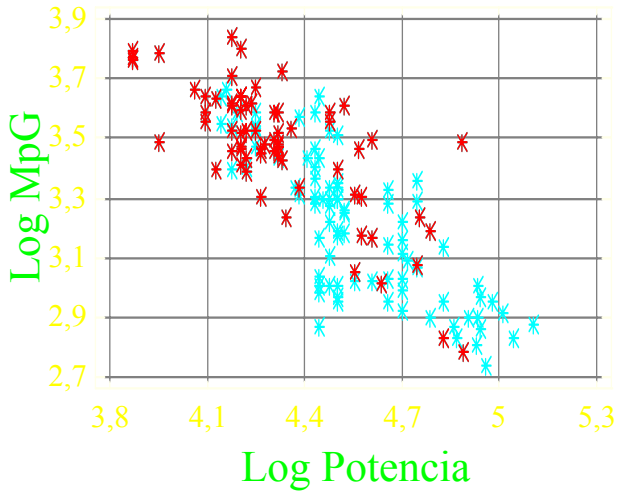


Figura 14: Gráfico de Consumo y Potencia en función del origen

Para comprobarlo realizamos la regresión entre las variables Log MpG, Log Potencia y la variable Origen que se define como:

$$Origen = \begin{cases} 1 \text{ Si el coche es de EE.UU} \\ 0 \text{ Si es Europeo o Japonés} \end{cases}$$

$$\begin{array}{llll} \log(MpG) = & 6.73 - & 0.75\log(Potencia) - & 0.09 Origen \\ \text{Estadístico } t & (29.77) & (-14.48) & (-3.26) \end{array}$$

$$R^2 = 0.69$$
$$\overline{R}^2 = 0.69$$

El modelo ajustado muestra claramente que existe un diferencia significativa entre el consumo de los coches norteamericanos y los de origen europeo o japonés. El estadístico $t = -3.26$ lo que indica que el origen es significativo. Así, los coches americanos circulan casi un 10% menos con un galón de gasolina que los europeos o los japoneses de la misma potencia. La figura 15 muestra los residuos del modelo. Puede observarse que no hay ninguna estructura, por lo que podemos concluir que el modelo es válido.

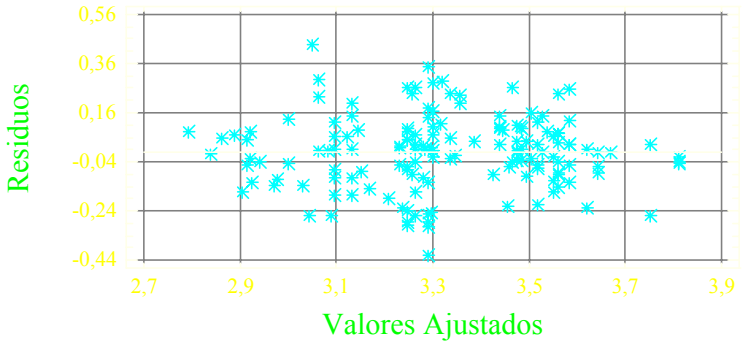


Figura 15: Gráfico de los residuos frente a valores ajustados del modelo 1.

14.2 Interacciones.

Hasta ahora hemos introducido variables dicotómicas para diferenciar grupos pero hemos exigido al modelo ajustar dos rectas paralelas. Indudablemente, nada asegura que la relación entre x e y sea tal que las rectas de regresión de ambos grupos sean así. Que las rectas sean paralelas implica pensar que a un incremento de x de un determinado valor, le corresponde un incremento de y del mismo valor en ambos grupos. La diferencia entre ambos grupos se mantiene constante aunque x aumente.

La realidad muchas veces no es así, y es preciso contrastar si las rectas son verdaderamente paralelas. Si las rectas no son paralelas se dice que existe una interacción entre la variable x y la variable z.

La modelización de la interacción es muy sencilla, pues el problema se resuelve realizando una regresión múltiple con y como variable dependiente y x, z y xz como variables independientes.

La variable xz es el producto de las variables x y z para cada observación. La regresión estimada será:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i + \hat{\beta}_3 x_i z_i \qquad (9)$$

Esta regresión estima dos rectas:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

cuando $z = 0$, y

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 + \hat{\beta}_3 x_i = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) x_i$$

cuando $z = 1$. Como puede observarse, cuando $z = 0$, la pendiente de la recta es $\hat{\beta}_1$. Mientras que cuando $z = 1$ la pendiente varía y es $(\hat{\beta}_1 + \hat{\beta}_3)$.

El método de análisis para saber si el grupo o característica cualitativa es significativa es muy sencillo. Se estima el modelo (9) y se contrasta si $\hat{\beta}_2$ y $\hat{\beta}_3$ son significativas.

Ejemplo:

Vamos a estudiar si existen dieferencias en la relación entre alturas y pesos por sexos. Para ello utilizaremos los datos de la figura 16, que representan los pesos y alturas de 117 estudiantes de la Universidad Politécnica de Madrid.

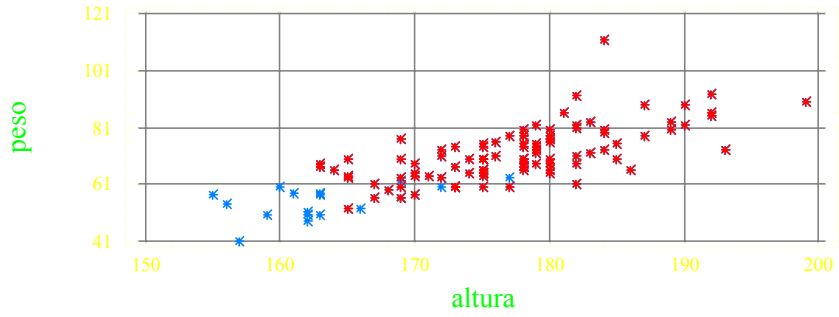


Figura 16: Alturas y Pesos por sexos. Datos de la figura 1.

Los puntos claros representan las mujeres y los oscuros los hombres. Se define la variable dicotómica Sexo como:

$$Sexo_i = \begin{cases} 0 \text{ si la persona } i \text{ es hombre} \\ 1 \text{ si la persona } i \text{ es mujer} \end{cases}$$

La regresión obtenida es:

Peso=	−77.79 +	0.84Altura −	5.18Sexo
Estadístico t		(9.29)	(−2.34)

$$\overline{R}^2 = 0.61$$

que indica que el Sexo influye en la relación entre Pesos y Alturas. Así, si una persona mide 1,70m de altura, su peso esperado es 65.01Kg si es un hombre. Si es una mujer, su peso será 65.01-5.18=59.83Kg. Vamos a estudiar si existe interación entre las variables Altura y Sexo. Para ello definimos una variable Altura*Sexo que toma los valores:

$$\text{Altura*Sexo}_i = \begin{cases} 0 \text{ si la persona } i \text{ es hombre} \\ \text{Altura}_i \text{ si la persona } i \text{ es mujer} \end{cases}$$

La regresión estimada indica que la variable Altura x Sexo no es significativa (Su t es menor que 2).

15. Variables Politómicas

En la sección anterior hemos estudiado la introducción de variables dicotómicas, es decir que podemos dividir la población en dos grupos. Sin embargo, en la práctica es habitual encontrar ejemplos en los que la población de estudio puede ser clasificada en varios grupos. Ejemplos de ésto pueden ser:

- Se modeliza la Renta que perciben diversos trabajadores en función de los estudios terminados. Los estudios pueden ser Primarios Medios o Superiores.
- Edad de Jubilación de los trabajadores en función del Sector de Actividad. El sector puede ser Industria, Agricultura, Construcción y Servicios.
- Cualquier modelización que dependa de Comunidad Autónoma (17 Comunidades en España)

La introducción de variables cualitativas que pueden tomar más de dos valores es muy simple. Para ilustrar el método utilizaremos un ejemplo.

Ejemplo:

Los datos que se van a analizar son las edades de jubilación (JUBI) de 394 varones de edad avanzada en 1987. Los datos proceden de la Encuesta de Población Activa (EPA) del segundo trimestre de 1987. Los hombres en cuestión se jubilaron durante los doce meses previos a la encuesta y se conoce por tanto su edad de jubilación. Las variable de interés son relativas a tres temas: Nivel de estudios del individuo, Sector de Actividad en que desarrollaba su trabajo, si estaba en paro antes de jubilarse y tipo de empleo que tenía el encuestado.

La modelización de las variables se realiza de la siguiente forma:

- Años de Estudios: Se modeliza a través de los años de estudio (AES). El modelo requiere introducir la variable AES y su cuadrado, AES^2.
- Sector de Actividad: Se definene tres sectores, Industria, Construcción y Servicios. Se definen tres variables dicotómicas una para cada sector. Así, la variable IND es una variable dicotómica que toma el valor 1 si el individuo trabajaba en la industria antes de jubilarse y el

valor cero en caso contrario. Es decir:

$$\text{IND}_i = \begin{cases} 1 \text{ si la persona } i \text{ trabajaba en la industria antes de jubilarse} \\ 0 \text{ si la persona } i \text{ no trabajaba en la industria antes de jubilarse} \end{cases}$$

$$\text{CONS}_i = \begin{cases} 1 \text{ si la persona } i \text{ trabajaba en la construcción antes de jubilarse} \\ 0 \text{ si la persona } i \text{ no trabajaba en la construcción antes de jubilarse} \end{cases}$$

$$\text{SERV}_i = \begin{cases} 1 \text{ si la persona } i \text{ trabajaba en los servicios antes de jubilarse} \\ 0 \text{ si la persona } i \text{ no trabajaba en los servicios antes de jubilarse} \end{cases}$$

- Tipo de Empleo. Existen tres posibilidades: Asalariado Público, Privado y Autónomo. Definimos una variable dicotómica para cada clase:

$$\text{ASALPUB}_i = \begin{cases} 1 \text{ si la persona } i \text{ era asalariado público antes de jubilarse} \\ 0 \text{ si la persona } i \text{ no era asalariado público antes de jubilarse} \end{cases}$$

$$\text{ASALPRI}_i = \begin{cases} 1 \text{ si la persona } i \text{ era asalariado privado antes de jubilarse} \\ 0 \text{ si la persona } i \text{ no era asalariado privado antes de jubilarse} \end{cases}$$

$$\text{AUTON}_i = \begin{cases} 1 \text{ si la persona } i \text{ era autónomo antes de jubilarse} \\ 0 \text{ si la persona } i \text{ no era autónomo antes de jubilarse} \end{cases}$$

Para estimar la regresión, se introducen para cada grupo de k variables dicotómicas $k - 1$ variables cualitativas. Es decir se deja fuera una de las variables. La razón de hacer ésto, es que si introducimos las k variables, el modelo tendría multicolinealidad exacta y no se podrían estimar los parámetros. La ecuación de regresión estimada es:

$$\begin{aligned} JUBI = \; & 65,49 \; -0.36AES \; +0.02AES^2 \\ & \qquad\quad (-2.27) \qquad (2.81) \end{aligned}$$

$$\begin{aligned} & -1.83ASALPUB \; -1.88ASALPRI \\ & (-4.23) \qquad\quad (-4.03) \end{aligned}$$

$$\begin{aligned} & -0.97IND \; -.66CONS \; -3.39PARADO \\ & (-2.51) \qquad (-1.13) \qquad (-4.73) \end{aligned}$$

La ecuación muestra que las variables dicotómicas son significativas. Utilizando la ecuación estimada, podemos obtener la edad estimada de jubilación de un trabajador de los grupos excluidos, es decir de un varón autónomo, que trabajaba en el sector servicios y no estaba parado. Este individuo tiene un valor 0 en las variables ASALPUB, ASALPRI, IND Y CONS. Por tanto su ecuación de regresión será:

$$\begin{aligned} JUBI = \; & 65,49 \; -0.36AES \; +0.02AES^2 \; -3.39PARADO \\ & \qquad\quad (-2.27) \qquad (2.81) \qquad\quad (-4.73) \end{aligned}$$

Así, por ejemplo, si la persona tiene 18 años de estudio (Estudios superiores) y no está parada, cabe esperar que se jubile a los

$$JUBI = 65,49 - 0.36x18 + 0.02x18^2 - 3.39x0 = 65.49\text{Años.}$$

Si la persona estuviera parada, se jubilaría 3.39 años antes.

Si la persona trabajase en el sector industrial, la ecuación de regresión nos dice que se jubilará 0.97 años antes que en el sector servicios. El sector de la construcción, como puede observarse no es significativo, por lo que la persona se jubilará a la misma edad si pertenence a la construcción que si pertenece a los servicios.

La modelización de variables politómicas es muy simple:

1. Se definen para cada variable politómica con k posibilidades, k variables dicotómicas que toman el valor 1 si la observación pertenece a la clase que estamos definiendo y 0 en caso contrario. Por ejemplo, si la variable es Comunidad Autónoma, definiremos 17 variables dicitómicas, que tomarán el valor 1 si la obsevación pertenece a la comunidad autónoma de que se trate y 0 en caso contrario.
2. Se introducen en el modelo $k - 1$ variables dicotómicas.
3. La interpretación se hace substituyendo los valores delas variables dummmies en la regresión estimada.