Regresión Lineal Múltiple

M.T. López, N.Corral

1. Introducción

Existen numerosas situaciones prácticas en las que se tienen observaciones de varias variables, y es razonable pensar en una relación entre ellas. En estos casos puede ser de gran interés la determinación, si existe, de esta relación expresada a través de una relación funcional ya que ello permitirá conocidos los valores de algunas variables, efectuar predicciones sobre los valores de otra, y también ancorrelalizar el tipo de relación de una variable con otras. El objetivo de la regresión será, por tanto, buscar un modelo funcional que nos permita explicar el comportamiento de una variable Y, *variable dependiente*, a través de los valores que toman otras variables $X_1, \ldots, X_p$ llamadas *variables independientes*. Es decir se buscaran relaciones del tipo $Y = f(X_1, \ldots, X_p) + \epsilon$.

Por simplicidad, nos centraremos por el momento en funciones $f(X_1, ..., X_p)$ lineales, obteniéndose así el modelo de regresión lineal.

Es importante señalar que el hecho de poner una variable Y como función de otras más una perturbación aleatoria no quiere decir que haya una relación de causalidad, sino que sólo se busca la existencia de una relación cuya forma y fuerza se quiere analizar.

Posteriormente se analizará que variables aportan explicación 'significativa' al comportamiento de la variable respuesta Para la realización de test de hipótesis serán necesarias condiciones sobre el comportamiento de la misma. Se estudiará también como comprobar las hipótesis sobre el modelo, para la correcta realización de las inferencias.

2. Regresión lineal

De acuerdo con lo ya se ha mencionado el objetivo de la regresión lineal múltiple es buscar el modelo lineal que mejor explique el comportamiento de una variable Y, en función de un conjunto de variables independientes y un término aleatorio ϵ que representa el error del modelo, es decir

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon = \vec{X}^t \vec{\beta} + \epsilon$$

siendo los elementos del modelo :

- $\vec{\beta}^t = (\beta_0, \ldots, \beta_p)$, parámetros del modelo, fijos y desconocidos.
- $X_1, \ldots, X_p$, variables correlindependientes o explicativas fijadas por el experimentador, también llamadas regresores.
- ϵ una variable aleatoria no observable.

Nota: Nada impide que los regresores sean transformaciones adecuadas de las variables originales. Por ejemplo, si pensamos que la variable aleatoria Y depende del cuadrado de X_j y de forma lineal de otras variables, podríamos especificar un modelo de regresión:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j^2 + \cdots + \beta_p X_p + \epsilon$$

Análogamente, si pensáramos que la variable aleatoria W se genera del siguiente modo: $W = kZ_1^{\beta_1} Z_2^{\beta_2} \nu$ siendo ν una perturbación aleatoria no negativa (por ejemplo, con distribución logarítmico normal), nada impediría que tomásemos logaritmos para obtener el modelo lineal $Y = log(W) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. En definitiva lo que se necesita es una expresión de la variable dependiente, o una transformación biyectiva, como función lineal de los parámetros.

Para estimar los parámetros del modelo se utiliza una muestra de n observaciones de la variable aleatoria Y, y los correspondientes valores de las p variables explicativas X_i, ello nos permitirá tener las siguientes n igualdades :

$$y_1 = \beta_0 + \beta_1 x_{1,1} + \cdots + \beta_p x_{1,p} + \epsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{2,1} + \cdots + \beta_p x_{2,p} + \epsilon_2$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_{n,1} + \cdots + \beta_p x_{n,p} + \epsilon\epsilon_k$$

En forma matricial, se escribirá

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x_1} + \cdots + \beta_p \vec{x_p} + \vec{\epsilon} = X\vec{\beta} + \vec{\epsilon}$$

siendo:

- $\vec{y}$ el vector de n de observaciones de la variable aleatoria respuesta o independiente Y.
- X la matriz, de dimensión $n \times (p+1)$, de valores de las variables explicativas en las n observaciones que se ha realizado; Se denomina matriz de regresión o de diseño, en general es de rango (p+1) aunque en algunos diseños experimentales puede tener rango menor.

- $\vec{\beta}$ el vector de parámetros de dimensión p+1.
- $\vec{\epsilon}$ el vector $n \times 1$ de valores de las perturbaciones aleatorias.

Debe hacerse notar que muchas veces no es posible fijar los valores de X, sino tan solo recoger datos de una muestra. Ello no afecta al desarrollo teórico que sigue, pero la inferencia sobre los parámetros es entonces condicional a los valores observados de X.

Para estimar $\vec{\beta}$ se tratará de minimizar los errores de predicción a través del criterio de mínimos cuadrados ordinario (MCO), para lo cual si $\widehat{\beta}$ denota el vector de estimadores de los parámetros, y $\vec{e}$ el vector de residuos, o valores de la variable residuo definida por $E = Y - \widehat{X}^t\widehat{\beta}$ (es decir, los residuos recogen la diferencia entre los valores muestrales observados de la variable Y y los ajustados a través del modelo), $\widehat{\beta}$ es el valor de parámetro que minimiza

$$g(\vec{\beta}) = \|\vec{y} - X\vec{\beta}\|^2 = (\vec{y} - X\vec{\beta})^t(\vec{y} - X\vec{\beta}) = \sum_{i=1}^{n}(y_i - x_i^t\vec{\beta})^2$$

o equivalentemente

$$g(\vec{\beta}) = \vec{y}^y\vec{y} - 2\vec{y}^tX\vec{\beta} + \vec{\beta}^tX^tX\beta$$

utilizando los procedimientos usuales de minimización se obtiene que los puntos críticos son las soluciones de

$$\frac{\partial g(\vec{\beta})}{\partial \vec{\beta}} = -2X^t\vec{y} + 2X^tX\vec{\beta} = \vec{0}$$

es decir, siempre que se pueda invertir la matriz $(X^tX)^{-1}$:

$$\widehat{\beta} = (X^tX)^{-1}X^t\vec{y}.$$

Con esta estimación del vector de parámetros se obtiene el valor de las predicciones que será

$$\widehat{y} = X\widehat{\beta} = X(X^tX)^{-1}X^t\vec{y} = P_X\vec{y}$$

siendo P_X la matriz de proyección ortogonal sobre el subespacio generado por las columnas de la matriz X

Nota: Si la matriz (X^tX) no tiene inversa o si presenta problemas en su inversión por tener un autovalor muy próximo a cero se debe a que la matriz X no es de rango completo y aparece un problema que se llama de multicolinealidad, existen variables explicativas que se pueden expresar linealmente en función de las otras.

La proyección ortogonal sobre el subespacio generado por las columnas de la matriz X siempre existe pero la forma de expresar esa proyección no es única y se produce un problema de inestabilidad de la estimación de los coeficientes.

Los residuos del modelo son, por tanto,

$$\vec{e} = \vec{y} - \widehat{y} = \vec{y} - P_X\vec{y} = (I - P_X)\vec{y} = Q_X\vec{y}$$

siendo Q_X la matriz de proyección sobre el subespacio ortogonal al generado por las columnas de la matriz X.

Para estudiar las propiedades del estimador $\widehat{\beta}$, que es un vector aleatorio dependiente del v.a. $\vec{y}$ que, a su vez, por el modelo $\vec{y} = X\vec{\beta} + \vec{\epsilon}$ y por ser la matriz X no aleatoria, depende del v.a. $\vec{\epsilon}$, requeriremos las siguientes condiciones:

- $E[\vec{\epsilon}] = \vec{0}$ (los residuos tienen media 0, si esto no fuese cierto y su media fuese $\mu\vec{1}$ se trabajaría con $\beta_0^* = \beta_0 + \mu$).
- $Var[\vec{\epsilon}] = E[\vec{\epsilon}^t\vec{\epsilon}] = \sigma^2 I$ (Los residuos son linealmente independientes y homocedásticos)
- $rango(X) = p + 1 < n$ para poder invertir la matriz X^tX.

Teorema 2.1. *Bajo las condiciones anteriores se verifican las siguientes propiedades del estimador* $\widehat{\beta}$

1. $E[\widehat{\beta}] = \vec{\beta}$ *(El estimador es centrado).*
2. $Var(\widehat{\beta}) = \Sigma_{\widehat{\beta}} = \sigma^2(X^tX)^{-1}$.
3. $s^2 = \frac{1}{n-(p+1)}\sum_{i=1}^n e_i^2$ *es un estimador insesgado de* σ^2.

Sugerencia
$e = (I - P_x)y = (I - Px)(X\beta + \epsilon) = (I - Px)\,\epsilon$

$E(e^t\,e) = E(tr(e\,e^t)) = tr(E(e\,e^t)) = tr((I - Px)Cov(\epsilon)(I - P_X))$

Teorema 2.2. *(Teorema de Gauss- Markov) Bajo las condiciones anteriores se verifican las siguientes propiedades del estimador* $\widehat{\beta}$ *Si* $\widehat{\beta}$ *es el estimador mínimo cuadrático ordinario de* $\vec{\beta}$*, cualquier otro estimador* $\widehat{\beta}^*$ *de* $\vec{\beta}$ *que sea lineal e insesgado tiene matriz de covarianzas con elementos diagonales no menores que los de* $\Sigma_{\widehat{\beta}}$*. Es decir* $\widehat{\beta}$ *es el estimador lineal insesgado cuyas componentes tienen la varianza más pequeña.*

La expresión $\sum_{i=1}^n e_i^2$ se suele llamar suma de cuadrados de los errores (SCE) y es la norma del vector de residuos al cuadrado, y se puede expresar $\|\vec{e}\|^2 =$

$\vec{y}^t(I - P_X)\vec{y}$. El estadístico SCE es muy útil, como se vio en el teorema anterior, para tener un estimador insesgado de la varianza de los errores, que viene dada por $\widehat{\sigma}^2 = \frac{SCE}{n-p-1}$.

3. Coeficiente de determinación

Hasta el momento hemos encontrado el estimador $\widehat{\beta}$ que minimiza la SCE, pero no se conoce cual es el grado de ajuste de las predicciones a los datos. Para resolver este problema se trata de buscar que parte de la variabilidad de los datos se puede explicar mediante las variables independientes.

Teorema 3.1. *Bajo las condiciones del modelo de regresión se verifica:*

1. *La media muestra de las predicciones coincide con la media muestral de los datos de la variable independiente* $\overline{Y} = \overline{\widehat{Y}}$.
2. *La media muestral de los residuales es cero.*

Teorema 3.2. *Bajo las condiciones del modelo de regresión se verifica:*

1. *SCT=SCR+SCE siendo*

$$SCT = \|\vec{y} - \overline{y}\vec{1}\|^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

$$SCR = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \qquad SCE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

2. *El coeficiente de determinación* R^2 *definido por*

$$R^2 = \frac{SCR}{SCT}$$

verifica que $0 < R^2 < 1$

3.

$$R^2 = 1 \quad \Leftrightarrow \quad e_i = 0 \quad \forall i = 1, \ldots, n$$
$$R^2 = 0 \quad \Leftrightarrow \quad \widehat{y}_i = \overline{y} \quad \forall i = 1, \ldots, n$$

Claramente cuantas más variables se incluyan en el modelo de regresión mejor se ajustarán las predicciones a los datos aunque algunas variables aporten 'poco' a la explicación de la variable independiente para paliar este efecto se define el *coeficiente de determinación corregido o ajustado* definido por

$$R_c^2 = 1 - \frac{\text{varianza residual}}{\text{varianza total}} = 1 - \frac{SCE/(n-p-1)}{SCT/(n-1)} = 1 - \frac{n-1}{n-p-1}\frac{SCE}{SCT}$$

4. Inferencias en el modelo de regresión

Para realizar inferencias de tipo estimación por intervalo y test de hipótesis suele imponerse, además de las condiciones previas, la de normalidad $\vec{\epsilon} \equiv N(\vec{0}, \sigma^2 I)$ lo que permitirá obtener distribuciones de los diferentes estadísticos, generalmente utilizando resultados de formas cuadráticas.

Teorema 4.1. *Dado $\vec{y} = X\vec{\beta} + \vec{\epsilon}$ con $X_{n,p+1}$ una matriz fija de rango $p+1$ y $\vec{\epsilon} \equiv N(\vec{0}, \sigma^2 I)$ se cumplen las siguientes propiedades*

1. $\widehat{\beta} \equiv N_{p+1}(\vec{\beta}, \sigma^2 (X^t X)^{-1})$ *(El estimador es centrado y normal).*
2. $\widehat{\beta}$ *es independiente de SCE y por tanto de* $\widehat{\sigma}^2$.
3. $\frac{SCE}{\sigma^2} \equiv \chi^2_{n-(p+1)}$.

Sugerencia: para la independencia de β y σ^2 estudiar la de β y e
Como β y e son funciones lineales de y que sigue una distribución normal, su independencia equivale a comprobar que su matriz de varianzas covarianzas es cero.

$$Cov(\beta, e) = Cov((X^t X)^{-1} X^t y; (I - P_X)y) = (X^t X)^{-1} X^t \, Cov(y) \, (I - P_X)$$

$$= \sigma^2 (X^t X)^{-1} X^t \, (I - P_X) = O$$

Con las distribuciones que hemos obtenido, bajo los supuestos de normalidad para los residuos, se pueden realizar distintos tipos de inferencia como

- Intervalos de confianza para $E(Y/\vec{x}_0)$

  Puesto que para el modelo anterior $E(Y/x_0) = \vec{x}_0^t \beta$ un estimador insesgado será

  $$\vec{x}_0^t \widehat{\beta} \equiv N_1(\vec{x}_0^t \beta, \sigma^2 \vec{x}_0^t (X^t X)^{-1} \vec{x}_0)$$

  y utilizando que $\widehat{\beta}$ es independiente de $\widehat{\sigma}^2$ se puede construir la siguiente función pivote

  $$\frac{\vec{x}_0^t \widehat{\beta} - \vec{x}_0^t \beta}{\widehat{\sigma} \sqrt{\vec{x}_0^t (X^t X)^{-1} \vec{x}_0}} \equiv t_{n-p-1}$$

  de donde el intervalo de confianza con coeficiente $(1-\alpha)$ será

  $$\left(\vec{x}_0^t \widehat{\beta} \pm F^{-1}_{t_{n-p-1}} (1 - \alpha/2) \widehat{\sigma} \sqrt{\vec{x}_0^t (X^t X)^{-1} \vec{x}_0} \right)$$

- Intervalo de confianza para la predicción de $Y/\vec{x}_0$ Bajo la hipótesis de que se trata de un individuo genérico de (independiente con la muestra) se tendrá que:

  $$Y/\vec{x}_0 - \vec{x}_0^t\widehat{\beta} \equiv N_1\left(0, \sigma^2(\vec{x}_0^t(X^tX)^{-1}\vec{x}_0 + 1)\right)$$

  que razonando como antes conducirá al intervalo, con coeficiente $(1-\alpha)$,

  $$\left(\vec{x}_0^t\widehat{\beta} \pm F^{-1}_{t_{n-p-1}}(1-\alpha/2)\widehat{\sigma}\sqrt{\vec{x}_0^t(X^tX)^{-1}\vec{x}_0 + 1}\right)$$

- Intervalos de confianza para combinaciones lineales $\vec{h}^t\vec{\beta}$

  se utilizará la función pivote

  $$\frac{\vec{h}^t\widehat{\beta} - \vec{h}^t\vec{\beta}}{\widehat{\sigma}\sqrt{\vec{h}^t(X^tX)^{-1}\vec{h}}} \equiv t_{n-p-1}$$

  como caso particular se tienen los intervalos de confianza para cada una de las componentes del vector de parámetros.

Todos estos intervalos son para parámetros unidimensionales también se pueden calcular regiones de confianza para $\phi = A\beta$ si $rango(A_{q,p+1}) = q$ que se basarían en su estimador

$$\widehat{\phi} = A\widehat{\beta} \equiv N_q\left(\phi, \sigma^2A(X^tX)^{-1}A^t\right)$$

lo que permite construir el estadístico

$$(\widehat{\phi} - \phi)^t(\sigma^2A(X^tX)^{-1}A^t)^{-1}(\widehat{\phi} - \phi) \equiv \chi^2_q$$

y por la independencia de $\widehat{\beta}$ con la suma de cuadrados del error se tendrá

$$\frac{(\widehat{\phi} - \phi)^t(A(X^tX)^{-1}A^t)^{-1}(\widehat{\phi} - \phi)}{qSCE/(n-p-1)} \equiv F_{q,n-p-1}$$

a partir del cual se pueden construir las regiones de confianza

4.1. Estimación con restricciones

En ocasiones se imponen restricciones al parámetro β para que el modelo sea coherente con la realidad, o para evitar problemas de multiples soluciones cuando hay problemas en la matriz de diseño. Nos restringiremos a restricciones

que pueden ser formuladas de la forma lineal $A\vec{\beta} = \vec{c}$. Es decir se planteará el problema de

$$min\|\vec{y} - X\vec{\beta}\| \quad \text{sujeto a} \quad A\vec{\beta} = \vec{c}$$

que se puede abordar por dos procedimientos diferentes, el método de Lagrange o a través de proyecciones ortogonales.

Para utilizar el método de las proyecciones se necesita formular un modelo equivalente en el que no aparezca el $\vec{c}$, por ello expresamos la restricción inicial $A\vec{\beta} = \vec{c}$ como $A(\vec{\beta} - \vec{\beta}_0) = \vec{0}$ (tiene que existir al menos una solución $\vec{\beta}_0$ ya que en caso contrario la restricción no tendría de sentido), se considera el parámetro $\vec{\gamma} = \vec{\beta} - \vec{\beta}_0$ que transforma la restricción en $A\vec{\gamma} = \vec{0}$ y el modelo de regresión en

$$\vec{y} = X\vec{\beta} + \vec{\epsilon} \Leftrightarrow \vec{y} - X\vec{\beta}_0 = X(\vec{\beta} - \vec{\beta}_0) + \vec{\epsilon} \Leftrightarrow \ \vec{y}^* = X\vec{\gamma} + \vec{\epsilon}$$

con lo que el problema inicial se transforma en

$$min\|\vec{y}^* - X\vec{\gamma}\| \quad \text{sujeto a} \quad A\vec{\gamma} = \vec{0}$$

La solución a este modelo trata de buscar $\vec{\theta} = X\vec{\gamma} \in M(X)$; como $X^t\vec{\theta} = X^tX\vec{\gamma}$ se tiene que cumplir que $\vec{\gamma} = (X^tX)^{-1}X^t\vec{\theta}$ y por la restricción $A\vec{\gamma} = A(X^tX)^{-1}X^t\vec{\theta} = \vec{0}$, es decir $\vec{\theta} \in M(X) \cap N(A(X^tX)^{-1}X^t)$ es decir $\widehat{\theta} = X\widehat{\gamma}$ va a ser la proyección de $\vec{y}^*$ sobre el subespacio $W = M(X) \cap N(A(X^tX)^{-1}X^t)$, de donde se obtendrá que el parámetro estimado es:

$$\widehat{\beta}_h = \widehat{\beta} - (X^tX)^{-1}A^t[A(X^tX)^{-1}A^t]^{-1}(A\widehat{\beta} - \vec{c})$$

Método de los multiplicadores de Lagrange para la condición $A\beta = c$

$$g(\beta, \lambda) = y^ty - 2\beta^tX^ty + X^tX\beta + \lambda^t(A\beta - c)$$
$$\frac{d\,g}{d\,\beta} = -2X^ty + 2(X^tX)\beta + A^t\lambda =$$
$$\frac{d\,g}{d\,\lambda} = A\beta - c = 0$$

4.2. Contraste de hipótesis lineales

Los contrastes de la forma $H_0: A\vec{\beta} = \vec{c}$ con $rango(A_{q,p+1}) = q$ (esta formulación engloba la mayoría de los test usuales sobre los coeficientes, como: valores determinados, comparación de los mismos, nulidad de componentes) se realizan a través del siguiente resultado

Teorema 4.2. *Dado $\vec{y} = X\vec{\beta} + \vec{\epsilon}$ con $X_{n,p+1}$ una matriz fija de rango $p+1$ y $\vec{\epsilon} \equiv N(\vec{0}, \sigma^2 I)$ y una matriz $A_{q,p+1}$ de rango q. Supongamos que $\widehat{\beta}_h$ es el estimador mínimo cuadrático de $\vec{\beta}$ bajo la restricción $A\vec{\beta} = \vec{c}$ se verifica que:*

1. *Si $SCE_h = (\vec{y} - X\widehat{\beta}_h)^t(\vec{y} - X\widehat{\beta}_h)$ entonces*

$$SCE_h - SCE = (A\widehat{\beta} - \vec{c})^t(A(X^tX)^{-1}A^t)^{-1}(A\widehat{\beta} - \vec{c})$$

2. *$E(SCE_h - SCE) = q\sigma^2 + (A\vec{\beta} - \vec{c})^t(A(X^tX)^{-1}A^t)^{-1}(A\vec{\beta} - \vec{c})$*

3. *Si la hipótesis $H_0 : A\vec{\beta} = \vec{c}$ es cierta entonces*

$$F = \frac{\frac{SCE_h - SCE}{q}}{\frac{SCE}{n-p-1}} = \frac{(A\widehat{\beta} - \vec{c})^t(A(X^tX)^{-1}A^t)^{-1}(A\widehat{\beta} - \vec{c})/q}{SCE/(n-p-1)} \equiv_{H_0} F_{q,n-p-1}$$

4. *Si $\vec{c} = \vec{0}$ el estadístico F anterior es*

$$F = \frac{n-p-1}{q}\frac{\vec{y}^t(P - P_h)\vec{y}}{\vec{y}^t(I - P)\vec{y}}$$

con P_h la matriz de proyección sobre el subespacio $W = M(X) \cap N(A(X^tX)^{-1}X)$

4.3. Contraste de la regresión o $R^2 = 0$

Uno de los contrastes más usuales en todos los análisis de regresión es lo que se conoce como el contraste de la regresión, o anova de la regresión, que plantea como hipótesis nula la no influencia de las variables explicativas en el respuesta, lo que se formula como $H_0 : \beta_1 = \cdots = \beta_p = 0$ o lo que es equivalente $H_0 : \rho^2 = 0$ siendo ρ^2 el mayor coeficiente de correlación lineal entre Y y una combinación lineal de las X_i.

Esta hipótesis se puede formular de forma lineal $A\vec{\beta} = \vec{0}$ con $A = (\vec{0}, I_p)$ en este caso $SCE_h = SCT$ y, por tanto, el estadístico F es

$$F = \frac{\frac{SCR}{p}}{\frac{SCE}{n-p-1}} = \frac{n-p-1}{p}\frac{SCT - SCE}{SCE} = \frac{n-p-1}{p}\frac{R^2}{1-R^2} \equiv_{H_0} F_{p,n-p-1}$$

5. Construcción de modelo de regresión

Hasta ahora hemos dado por supuesto que el modelo lineal que se estima es el 'correcto', es decir, que la variable aleatoria Y efectivamente se genera de la manera: $Y = \beta_0 + \beta_1 X_1 \cdots + \beta_p X_p + \epsilon$, pero en la práctica no suele tenerse

un conocimiento exacto del modelo y pueden considerarse variables que no se deberían estar en el modelo, ya que su aportación a la explicación de la variable independiente no es significativo. Surge así el problema de la regresión por etapas donde se van incluyendo, en pasos sucesivos variables que aporten explicación de la variable independiente y se eliminan aquellas que no aportan información. Otro problema que se plantea en la construcción del modelo es el tratamiento de las variables independientes cuando estas son cualitativas, ello se suele realizar mediante la construcción de variables ficticias o dummy que reflejan cada una de las modalidades de la variable respuesta.

5.1. Regresión por etapas

Dadas las datos de las variables independientes $(\vec{1}, \vec{x}_1, \ldots, \vec{x}_p)$ y la variable independiente $\vec{y}$ se sigue el siguiente procedimiento:

Etapa 0 Se incluye el $\vec{1}$ (es decir se centran los datos)

Predicciones: $\widehat{y} = P_{\vec{1}}\vec{y} = \overline{y}\vec{1}$
Residuales: $\vec{e} = (I - P_1)\vec{y} = Q_{\vec{1}}\vec{y} = \vec{y} - \overline{y}\vec{1}$
Varianza residual: $SCE_0 = SCT = \vec{y}^t(I - P_{\vec{1}})\vec{y} = \sum(y_i - \overline{y})^2$

Etapa 1 Se incluye la variable independiente X_i que mejor explica los residuales de la etapa anterior $Q_{\vec{1}}\vec{y}$

Posible aportación nueva de cada variable $Q_{\vec{1}}\vec{x}_i$ (lo otro ya está explicado por el modelo anterior)
Coeficiente de regresión de cada variable: $\widehat{\beta}_i = (\vec{x}_i^t Q_{\vec{1}}\vec{x}_i)^{-1}\vec{x}_i^t Q_{\vec{1}}\vec{y}$
Predicción con la variable i : $\widehat{y}_1 = Q_{\vec{1}}\vec{x}_i\widehat{\beta}_i$
Residual con la variable i: $\vec{e}_1 = Q_{\vec{1}}\vec{y} - \widehat{y}_1 = Q_{\vec{1}}(\vec{y} - \vec{x}_i\widehat{\beta}_i)$
Variación explicada con la variable i: $SCR_1 = \widehat{y}_1^t\widehat{y}_1$
Variación residual con la variable i: $SCE_1 = \vec{e}_1^t\vec{e}_1$
Proporción de variación explicada por la variable i:

$$R_1^2 = \frac{SCR_1}{SCR_1 + SCE_1} = \frac{SCR_1}{SCT}$$

se coge la variable i que tenga mayor R_1^2 si es que aporta 'algo' a la explicación de la variable independiente (se comprueba si $\beta_i = 0$

Etapa k+1 Se utiliza el siguiente resultado, más general, para saber que variable se introduce en la etapa k+1, que será la variable, no en el modelo, que más aporte a la explicación de los residuales de la etapa anterior.

Teorema 5.1. *Dado un modelo lineal* $\vec{y} = W\vec{\delta} + \vec{\epsilon}$ *donde la matriz de diseño,* $W_{n,p+q}$ *se puede dividir en dos submatrices* $W = (X, Z)$ *con* $X_{n,p}$, $Z_{n,q}$ *matrices de rangos p,q, respectivamente, es decir* $\vec{y} = X\vec{\beta} + Z\vec{\gamma} + \vec{\epsilon}$. *Si* $\widehat{\delta}$ *es el estimador de mínimos cuadrados de* δ *con con componentes* $\widehat{\delta} = (\widehat{\beta}_g, \widehat{\gamma}_g)$ *entonces se verifican las siguientes propiedades:*

1. $\widehat{\gamma}_g = (Z^t Q_X Z)^{-1} Z^t Q_X \vec{y}$ *(en las columnas de Z y en* $\vec{y}$ *se elimina todo lo que se puede explicar por X y se hace la regresión)*
2. $\widehat{\beta}_g = (X^t X)^{-1} X^t (\vec{y} - Z\widehat{\gamma}_g) = \widehat{\beta} - (X^t X)^{-1} X^t Z \widehat{\gamma}_g$
3. $SCE_g = (\vec{y} - Z\widehat{\gamma}_g)^t Q_X (\vec{y} - Z\widehat{\gamma}_g) = \vec{y}^t Q_X \vec{y} - \widehat{\gamma}_g^t Z^t Q_X \vec{y}$
   El vector de residuos global es el vector de residuos de X corregido por lo explicado por $Q_X Z$
4. *Se consideran las matrices* $L = (X^t X)^{-1} X^t Z$, $M = (Z^t Q_X Z)^{-1}$ *la matriz de varianzas -covarianzas de* $\widehat{\delta}$ *es*

$$\sigma^2 \begin{pmatrix} (X^t X)^{-1} + LML^t & -LM \\ -ML^t & M \end{pmatrix}$$

El caso más general cuando se quiere construir un modelo que permita predecir el comportamiento de cierta variable es que solo tengamos una idea aproximada de que variables se deben incluir como predictoras. Por ello es interesante disponer de de criterios que permitan comparar entre diferentes alternativas.
Por supuesto una vez realizada esta selección queda a criterio del investigador decidir si se debe modificar, excluyendo algunas, realizando transformaciones o incluyendo nuevas variables.

Los criterios más habituales de comparación de modelos son los siguientes:

- No incremento del coeficiente de determinación ajustado (ya que el R^2 siempre aumenta.
- El criterio de Akaike , AIC, (Akaike Information Criterion) que consiste en considerar el valor de

$$AIC = n * log(\widehat{\sigma}_p^2) + 2p$$

  y elegir el que tenga valor mínimo, siendo p el numero de parámetros en el modelo y $\widehat{\sigma}_p^2$ la estimación MV correspondiente de la varianza de los errores,

al aumentar el número de variables disminuye la estimación pero aumenta 2p con lo que el criterio tiene en cuenta ambas cosas.

- BIC (Bayesian Information Criterion) que penaliza más los modelos de dimensión elevada. Se base en

$$BIC = n * log(\widehat{\sigma}_p^2) + p\, log(n)$$

- Criterio C_p de Mallows definido por

$$C_p = \sum_{i=1}^{n} (E(\widehat{y}_i - E(y_i))^2/\sigma^2 = p + (n-p)\frac{\widehat{s}_R^2(p) - \widehat{s}_R^2(k+1)}{\widehat{s}_R^2(k+1)}$$

y elegir el modelo con menor C_p

Los métodos habituales de regresión por etapas son:

- Hacia adelante (forward). Se empieza con un modelo sin ninguna variable predictora y se van incluyendo una a una hasta alcanzar un criterio de parada. El más habitual es que el coeficiente de la variable no sea significativamente diferente de cero.

- Hacia atráse (backward). Se empieza con un modelo en el que intervienen todas las variables predictoras y se van eliminando las que no tengan una aportación relevante en la predicción. El más habitual es que el coeficiente de la variable no sea significativamente diferente de cero

- Por pasos sucesivos (Stepwise) Es una combinación de los otros dos. Se va hacia adelante pero en cada etpa se comprueba si alguna de las variables seleccionadas deja de tener interés.

Definición 5.1. *La correlación parcial entre dos variables, por ejemplo $X_1 y X_2$ dadas otro conjunto de ellas $X = (1, X_3, \ldots, X_p)$ es la correlación entre los residuos de X_1 no explicados por X, $e_1 = X_1 - X\widehat{\beta}_1$, y los de X_2 no explicados por X, $e_2 = X_2 - X\widehat{\beta}_2$. se suele denotar por $R_{12,3\ldots p}$*

En el caso de tres variables se tiene

$$R_{12,3} = \frac{R_{12} - R_{12}R_{13}}{\sqrt{((1 - R_{13}^2)(1 - R_{23}^2)}}$$

y en el caso general se tiene que está relacionado con el coeficiente de determinación:

$$1 - R^2 = \frac{SCE}{SCT} = \frac{SCEy.x_1}{SCT}\frac{SCEy.x_1x_2}{SCEy.x_1}\cdots\frac{SCE_{y.x_1,\ldots,x_{p-1}}}{SCE_{y.x_1,\ldots,x_{p-2}}}\frac{SCE_{y.x_1,\ldots,x_p}}{SCE_{y.x_1,\ldots,x_{p-1}}}$$

es decir:

$$1 - R^2 = (1 - R^2_{yx_1})(1 - R^2_{yx_2.x_1})\ldots(1 - R^2_{yx_p.x_1\ldots x_{p-1}})$$

También se verifica que si R^2_k es el coeficiente de determinación de Y con las variables $(1, X_3, \ldots, X_k)$ y R^2_{k-1} es el coeficiente de determinación de Y con las variables $(1, X_3, \ldots, X_{k-1})$ entonces:

$$1 - R^2_k = \frac{SCE_g}{SCT} = \frac{SCE_{y.x_1,\ldots,x_{k-1}} - SCR_{yx_k.x_1,\ldots,x_{k-1}}}{SCT}$$

$$= (1 - R^2_{k-1})(1 - \frac{SCR_{yx_k.x_1,\ldots,x_{k-1}}}{SCR_{y.x_1,\ldots,x_{k-1}}}) = (1 - R^2_{k-1})(1 - R^2_{yx_k\,.x_1,\ldots,x_{k-1}})$$

de donde

$$R^2_k - R^2_{k-1} = (1 - R^2_{k-1})R^2_{yx_k.x_1,\ldots,x_{k-1}}$$

6. Análisis de residuales

Una vez estimado el modelo de regresión los residuos son un elemento importante para comprobar las hipótesis de linealidad, homocedasticidad y normalidad que has sido utilizadas en la inferencia sobre el modelo. Los residuos están definidos como $\vec{e} = (I - P_X)\vec{y} = Q_X\vec{y}$ y sabemos que están relacionados con la variable error ϵ , más concretamente con una muestra de esta variable,$\vec{\epsilon}$, a la que hemos impuesto $E(\vec{\epsilon}) = \vec{0},\ var(\vec{\epsilon}) = \sigma^2 I$.

Sin embargo si $\vec{1} \in X$ entonces $E(\vec{e}) = \vec{0}$ pero su matriz de covarianza verifica $var(\vec{e}) = var(Q_X\vec{y}) = Q_X\sigma^2 I Q^t_X = \sigma^2 Q_X = \sigma^2(I - P_X)$ es decir los residuos tienen media cero pero no son, en general, ni linealmente independientes, ni homocedásticos.

Un residuo grande indica un alejamiento del dato al modelo proporcionado, pero para saber cuando es grande se necesita conocer la varianza del mismo. La varianza de cada residuo es: $var(e_i) = \sigma^2(1 - p_{ii}) = \sigma^2(1 - x^t_i(X^tX)^{-1}x_i)$. Si la varianza es pequeña el modelo de regresión pasa por cerca del punto y si la varianza es grande el punto está alejado del modelo.

Debido a esta homocedasticidad los residuos se suelen studentizar o estandarizar (convertir a variable con igual varianza). El residuo estandarizado es:

$r_i = \frac{e_i}{\sqrt{\widehat{\sigma}^2(1-p_{ii})}}$ trata de eliminar el efecto de la heterocedasticidad (la distribución no tiene por que ser una t ya que no está garantizada la independencia del numerador con la SCE). Otra forma de estandarizar es estimar σ con el resto de los puntos para tener independencia del denominador y numerador y conseguir una distribución tipo t, obteniéndose los residuos estudentizados $e*_i = \frac{e_i}{\sqrt{\widehat{\sigma}^2_{-i}(1-p_{ii})}}$ que seguirá una distribución $t_{n-1-p-1}$.

Los puntos con un alto, o bajo, residuo estandarizado se llaman *outlier* Entre los outliers es importante detectar los *puntos de influencia* que son aquellos cuya presencia hacia que cambie mucho los valores de los coeficientes de regresión, para su detección se pueden seguir diversos criterios: buscar el residuo estandarizado cuando el punto no es utilizado para construir el modelo, comparar los coeficientes del modelo con y sin el punto $\widehat{\beta} - \widehat{\beta}_{-i}$, etc.

Entre las medidas de influencia más habituales se encuentran las siguientes:

1. Puntuaciones hat p_{ii}, que son los elementos de la diagonal principal de la matriz $X(X^tX)^{-1}X^t$.

2. Distancia de Cook

$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}_{-i})^t(X^tX)(\widehat{\beta} - \widehat{\beta}_{-i})}{p\widehat{\sigma}^2} = \frac{(\widehat{y} - \widehat{y}_{-i})^t(\widehat{y} - \widehat{y}_{-i})}{p\widehat{\sigma}^2}$$

3. $DFFITS$

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i-i}}{\hat{s}_{-i}\sqrt{p_{ii}}}; \quad |DFFIT_i| > 2\sqrt{\frac{p}{N}}$$

4. $DFBETAS$

$$DFBETAS_{ji} = \frac{\hat{\beta}_j - \hat{\beta}_{j-i}}{\hat{s}\sqrt{(X^tX)^{-1}_{jj}}}; \quad |DFBETA_{ij}| > \frac{2}{\sqrt{N}}$$

Sugerencias para los cálculos eliminando la observación i-ésima

1.

Teorema 6.1. *(Shermann-Morrison-Woodbury)*

$(A - zz^t)^{-1} = A^{-1} + A^{-1}\ zz^t\, A^{-1}/(1 - z^t\, A^{-1}\, z)$

2.
$$X^t X = \sum_{k=1}^{n} \vec{x}_k \, \vec{x}_k^t = \sum_{k \neq i} \vec{x}_k \, \vec{x}_k^t + \vec{x}_i \, \vec{x}_i^t = X_{-i}^t X_{-i} + \vec{x}_i \, \vec{x}_i^t$$

3.
$$(X_{-i}^t X_{-i})^{-1} = (X^t X)^{-1} + (X^t X)^{-1} \, \vec{x}_i \, \vec{x}_i^t \, (X^t X)^{-1} / (1 - p_{ii})$$

4.
$$X^t \, \vec{y} = \sum_{k \neq i} \vec{x}_k \, y_k + \vec{x}_i \, y_i = X_{-i}^t \, y_{-i} + \vec{x}_i \, y_i = X^t \vec{y} - \vec{x}_i \, y_i$$

5.
$$\hat{\beta}_{-i} = (X_{-i}^t X_{-i})^{-1} X_{-i}^t \, \vec{y}_{-i} = \hat{\beta} - (X^t X)^{-1} \vec{x}_i \, \vec{e}_i / (1 - p_{ii})$$

6.
$$e_{-i} = y_i - \vec{x}_i^t \, \hat{\beta}_{-i} = y_i - \vec{x}_i^t \, \hat{\beta} + \vec{x}_i^t \, (X^t X)^{-1} \, \vec{x}_i \, e_i / (1 - p_{ii})$$

6.1. Gráficos de residuos

Es usual realizar distintos gráficos para detectar la presencia de outliers, o la no adecuación del modelo a los datos. Los tipos de gráficos más utilizados son:

- Predicciones frente a residuales $(\widehat{y}_i, e_i)$. De acuerdo con las condiciones teóricas usuales, ambas variables son independientes y por lo tanto no se debe apreciar ninguna estructura en el gráfico
- Índice de cada observación frente a los residuales. Tienen interés cuando las observaciones son secuenciales en el tiempo, ya que permite detectar pautas de comportamiento como agrupamiento de residuos.
- Gráficos de normalidad de los residuales.
- Residuales frente a las variables incluidas en el modelo. Pueden aportar información sobre la forma de intervenir la variable en el modelo, que puede ser no lineal sino cuadrática, exponencial o logarítmica etc.
- Residuales frente a las variables excluidas del modelo.
- Gráficos parciales de residuos. Son del mismo tipo que las dos anteriores pero eliminando la influencia del resto de variables.
    1. La variable $\vec{x}_k$ frente a los residuales de la regresión de $\vec{y}$ con X_{-k}
    2. Los residuales de $\vec{x}_k$ con a X_k frente a los residuales de $\vec{y}$ con X_{-k}

7. Apéndice : Proyecciones ortogonales

Propiedades de las proyecciones ortogonales

- Si Ω es un subespacio vectorial de $\mathbb{R}^n$ entonces todo vector $y \in \mathbb{R}^n$ se puede descomponer de forma única como $y = u + v,\ u \in \Omega, v \in \Omega^{\perp}$
- Si Ω es un subespacio vectorial de $\mathbb{R}^n$ existe una única matriz de proyección sobre el P_Ω de modo que $P_\Omega y = u$, esta matriz se puede construir a través de una matriz X cuyas columnas son un sistema de generadores de subespacio de la forma siguiente $P_\Omega = X(X^tX)^{-1}X^t$
- Si P_Ω es la matriz de proyección sobre Ω entonces $Q_\Omega = I - P_\Omega$ es la matriz de proyección sobre $\Omega^{\perp}$
- P_Ω y Q_Ω son matrices simétricas e idempotentes
- Dados dos subespacios de $\mathbb{R}^n$, Ω_1 y Ω_2 se verifica que $(\Omega_1 \cap \Omega_2)^{\perp} = \Omega_1^{\perp} + \Omega_2^{\perp}$
- Dados dos subespacios de $\mathbb{R}^n$, Ω, ω tales que $\omega \subset \Omega$ entonces $P_\Omega P_\omega = P_\omega P_\Omega = P_\omega$
- Dados dos subespacios de $\mathbb{R}^n$, Ω, ω tales que $\omega \subset \Omega$ entonces $P_\Omega - P_\omega = P_{\Omega \cap \omega^{\perp}}$
- Sea A una matriz y dos subespacios de $\mathbb{R}^n$, Ω, ω tales que $\omega = N(A) \cap \Omega$ entonces $\Omega \cap \omega^{\perp} = M(P_\Omega A^t)$
- Sea A una matriz qxn rango(A)=q. Se verifica que $rango(P_\Omega A^t) = q$ si y solo si $M(A^t) \cap \Omega^{\perp} = 0$