Estadística II

Laura M. Castro Souto

Segundo Cuatrimestre
Curso 2000/2001

Modelos de Regresión

Diferencias con el Diseño de Experimentos

- Los **modelos de regresión** estudian relaciones numéricas entre variables *cuantitativas*, mientras que en diseño de experimentos las variables explicativas (factores) son de carácter *cualitativo*.
- El **diseño de experimentos** se usa básicamente para experimentar, esto es, una vez que se plantea el problema y se determinan los factores que pueden influir, se diseña el experimento y se observan los resultados. En cambio, con los modelos de regresión es habitual observar el comportamiento de varias variables para luego buscar relaciones funcionales entre las variables a partir de la muestra de observaciones multivariantes observadas (*regresión en diseño aleatorio*), aunque también se pueden fijar valores en las variables regresoras y experimentar para observar el comportamiento de la variable de interés (*regresión en diseño fijo*).

Cuando estudiamos la relación entre una variable de interés, **variable respuesta** (Y) y un conjunto de variables explicativas, **variables regresoras** (X_1, X_2,. . . ,X_k) puede ocurrir:

- Que exista una relación funcional entre ellas, en el sentido de que el conocimiento de las variables regresoras determine completamente el valor que toma la variable respuesta, esto es:

  $$Y = f(X_1, X_2, \ldots, X_k)$$

  Por ejemplo, la distancia recorrida por un móvil que se mueve a velocidad constante.

- Que no exista ninguna relación entre la variable respuesta y las variables regresoras, en el sentido de que el conocimiento

de éstas no proporcione ninguna información sobre el comportamiento de la otra:

$$Y = \varepsilon$$

Ejemplo: el dinero que gana una persona alta. Variables que parece que tienen relación pero no la tienen en realidad se involucran en *relaciones espúreas*.

- El caso intermedio, que exista una **relación estocástica** entre la variable respuesta y las variables regresoras, en el sentido de que el conocimiento de éstas nos permita predecir con mayor o menor exactitud el valor de la variable respuesta. Siguen por tanto un modelo:

$$Y = f(X_1, X_2, \ldots, X_k) + \varepsilon$$

Estas últimas son las relaciones que ocurren en la mayoría de las situaciones y que determinamos **modelos de regresión**. El objetivo es determinar la función f y el modelo probabilístico que sigue el error aleatorio ε[1].

Cuando se quiere estudiar la relación estocástica entre una variable de interés y un conjunto de variables regresoras se plantean diferentes problemas:

- ¿Qué variables explicativas se deben usar en el modelo? ¿Qué variables son *significativas*, es decir, su inclusión en el modelo mejora el conocimiento acerca del comportamiento de la variable de interés?

- ¿Qué función explica la relación entre la variable de interés y las variables explicativas o regresoras? ¿Es razonable suponer que la función es de una determinada *familia*, por ejemplo, lineal? En caso afirmativo, (**enfoque paramétrico lineal**, enfoque tradicional), el problema básico sería estimar

[1] ¿Cuál es la f? ¿Cuáles son las X's? ¿Cuánto y/o cómo influye cada X_i?

los parámetros de la familia supuesta a partir de la muestra. Es decir, si suponemos un modelo del tipo:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_i X_i + \varepsilon$$

el problema radica en estimar los parámetros α_0, α_1,. . . ,α_i y contrastar que la hipótesis supuesta es aceptable. Cuando no lo es, no sólo se suele cambiar de familia para seguir probando, sino que tambien se suelen transformar los datos, por ejemplo:

$$\begin{array}{rcl} Y & \rightarrow & \varphi \\ x & \rightarrow & \dfrac{1}{x} \end{array}$$

En este curso se estudian los **modelos de regresión lineal**; con estos modelos se trata de estudiar la *relación lineal* existente entre una variable respuesta Y y un conjunto de variables regresoras o explicativas X_1, X_2,. . . ,X_k. Para ello a partir de una muestra de observaciones: $\{(x_1, x_2, \ldots, x_k, y)\}_{i=1}^{n}$, se desea estudiar un modelo de regresión de la forma:

$$\boxed{Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_i X_i + \varepsilon}$$

Según la forma de recogida muestral, se distinguen *dos tipos de regresión*:

* Modelos de regresión con **diseño fijo**: las variables regresoras son *variables matemáticas predeterminadas* (elegidas por el experimentador). Este modelo se usa cuando se quiere conocer el comportamiento de la variable respuesta cuando las variables regresoras varían en una determinada dirección. En este caso se debe diseñar y realizar un experimento en el que las variables regresors se muevan en dicha dirección. Por tanto, con este

diseño se controla en todo momento el valor de las variables regresoras.

* Modelos de regresión con **diseño aleatorio**: las variables regresoras X_i son *variables aleatorias*. Este modelo se usa cuando se desea estudiar la relación entre la variable respuesta y las variables regresoras a partir de una muestra obtenida recogiendo los resultados de las variables en unidades de experimentación elegidas al azar. Esto es, el experimentador es un observador pasivo.

Las operaciones a realizar en ambos casos son las mismas, pero el desarrollo en diseño aleatorio es más complejo. Nosotros haremos diseño fijo.

Parámetros del Modelo

Estimación por mínimos cuadrados

Un primer objetivo en el estudio de este modelo es estimar los parámetros del mismo: α_0, α_1 y σ^2, a partir de las observaciones muestrales.

Una vez calculadas las estimaciones de los parámetros de la recta de regresión: $\hat{\alpha}_0$ y $\hat{\alpha}_1$, podemos calcular las **predicciones** para las observaciones muestrales, que vienen dadas por,

$$\hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i \qquad i = 1, 2, \ldots n$$

o, en forma matricial,

$$\hat{Y} = \hat{\alpha}_0 \vec{1} + \hat{\alpha}_1 \vec{X}$$

donde $\hat{Y}_i = (\hat{y}_1, \hat{y}_2, \ldots \hat{y}_n)$.

Denominamos **residuos** a

$$\vec{e} = \vec{Y} - \hat{Y}$$

es decir,

$$\textbf{Residuo}(e_i) = \text{Valor observado } (y_i) - \text{valor predicho } (\hat{y}_i)$$

Una vez obtenidas las **ecuaciones canónicas**, haciendo el desarrollo:

$$\left.\begin{array}{rcl} \bar{x}\bar{y} & = & \hat{\alpha}_0 \bar{x} + \hat{\alpha}_1 \bar{x}^2 \\ x\bar{y} & = & \hat{\alpha}_0 \bar{x} + \hat{\alpha}_1 \bar{x^2} \end{array}\right\} \qquad (2) - (1) \quad \Rightarrow$$

$$\bar{xy} - \bar{x}\bar{y} = \hat{\alpha}_1(\bar{x^2} - \bar{x}^2) \quad \Rightarrow \quad \mathbf{S_{xy} = \hat{\alpha}_1 S_x{}^2}$$

de donde se obtienen los estimadores mínimi-cuadráticos para los parámetros α_0 y α_1.

La estimación puede hacerse por dos métodos que son equivalentes bajo hipótesis de normalidad.

Estimación por máxima verosimilitud

Como $\dfrac{y_i}{x_i} \in N(\alpha_0 + \alpha_1 x_i, \sigma^2)$, la función de verosimilitud asociada a la muestra es

$$l(\alpha_0, \alpha_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(y_i - \alpha_0 - \alpha_1 x_i)^2\right)$$

de donde la función soporte es

$$L(\alpha_0, \alpha_1, \sigma^2) = -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i)^2$$

Maximizando esta función (se busca darles a α_0, α_1 y σ^2 los valores que hagan que la muestra obtenida sea la más probable) se obtienen los mismos estimadores que con mínimos cuadrados, ya que ignorando los dos primeros términos, maximizar

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i)^2$$

es como minimizar

$$\sum_{i=1}^{n}(y_i - \alpha_0 - \alpha_1 x_i)^2$$

que es ni más ni menos que la suma de residuos al cuadrado, que es lo que minimiza el otro método.

Los estimadores obtenidos son

$$\hat{\alpha}_{1,MV} = \frac{S_{XY}}{{S_X}^2}$$

siendo S_{XY} la covarianza muestral de X e Y, y ${S_X}^2$ la varianza muestral de X.

$$\hat{\alpha}_{0,MV} = \bar{y} - \alpha_{1,MV}\bar{x}$$

$${\hat{S}_{R,MV}}^2 = \frac{1}{n}\sum_{i=1}^{n} {e_i}^2 \quad \text{(estimador sesgado)}$$

Comentarios y propiedades

- La recta de regresión pasa por $(\bar{x}, \bar{y})$, que es el centro geométrico de la nube de datos.
- $\hat{\alpha}_1$ se denomina **coeficiente de regresión** y es la pendiente de la recta de regresión. Tiene una sencilla interpretación: nos indicará el crecimiento (o decrecimiento) de la variable respuesta Y asociada a un incremento unitario de la variable regresora X.
- La distribución de $\hat{\alpha}_1$ es una normal de media α_1 y varianza $\dfrac{\sigma^2}{n{S_X}^2}$, por lo tanto la varianza de $\hat{\alpha}_1$ verifica que:
  - disminuye al aumentar n
  - disminuye al aumentar ${S_X}^2$
  - aumenta al aumentar σ^2

En resumen, se verifica que:

$$\hat{\alpha}_1 \in N\left(\alpha_1, \frac{\sigma^2}{{S_X}^2 n}\right)$$

y es además un estimador insesgado. Para calcular intervalos de confianza:

$$\frac{\hat{\alpha}_1 - \alpha}{\frac{\sigma}{S_X}\sqrt{n}} \in N(0,1)$$

pero no podemos hacerlo porque desconocemos σ.

- El parámetro α_0, que indica la ordenada de la recta de regresión para $x = 0$, tiene menor importancia. La distribución de su estimador $\hat{\alpha}_0$ es una normal de media α_0 y varianza $\frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{n{S_X}^2} = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{{S_X}^2}\right)$, por lo tanto la varianza de $\hat{\alpha}_0$ verifica que:

  - disminuye al aumentar n
  - disminuye al aumentar ${S_X}^2$
  - aumenta al aumentar σ^2 (esto es, hasta aquí se comporta igual que $\hat{\alpha}_1$)
  - aumenta al aumentar $\bar{x}$

  En resumen, se verifica que

$$\hat{\alpha}_0 \in N(\alpha_0, \sigma\sqrt{\frac{1}{n}\left(1 + \frac{\bar{x}^2}{{S_X}^2})\right)}$$

- El estimador máximo verosímil de σ^2 es ${\hat{\sigma}_{MV}}^2$, cuya distribución es $\frac{n{\hat{\sigma}_{MV}}^2}{\sigma^2} \in {\chi_{n-2}}^2$, por tanto es un estimador sesgado,

$$E({\hat{\sigma}_{MV}}^2) = \frac{n-2}{n}\sigma^2$$

El número de grados de libertad es $n-2$ porque los n residuos verifican dos restricciones:

$$\sum_{i=1}^{n} e_i = 0$$
$$\sum_{i=1}^{n} e_i x_i = 0$$

Por este motivo se utiliza como estimador de σ^2 la varianza residual $\hat{S}_R{}^2$, dada por

$$\hat{S}_R{}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i{}^2$$

y cuya distribución es

$$\frac{(n-2)\hat{S}_R{}^2}{\sigma^2} \in \chi_{n-2}{}^2$$

A partir de este estadístico podemos obtener intervalos de confianza y test de hipótesis de la varianza poblacional σ^2.

Como ya hemos indicado, el parámetro α_0 tiene menor importancia y, en algunas situaciones, no tiene una interpretación realista si el cero no es un punto del rango de la X, por ejemplo, al estudiar la relación entre peso y altura de un colectivo de personas. Por ello, también tiene interés la ecuación de la recta de regresión dada en función del parámetro α_1, teniendo en cuenta:

$$\begin{aligned}\tilde{x} &= x - \bar{x}\\ \tilde{y} &= y - \bar{y}\\ \tilde{y} &= \hat{\alpha}_1 \tilde{x}\end{aligned}$$

La recta de regresión de X sobre Y es distinta de la recta de regresión de Y sobre X. En este caso tendremos que:

$$\hat{x}_i = \gamma_0 + \gamma_1 y_i$$

siendo

$$\hat{\gamma}_1 = \frac{S_{XY}}{{S_y}^2} \qquad \text{y} \qquad \hat{\gamma}_0 = \bar{x} - \hat{\gamma}_1 \bar{y}$$

Interpretación geométrica

Consideremos los siguientes vectores del espacio n-dimensional $\mathbb{R}^{\mathrm{n}}$:

$\vec{Y}$	$=$	$(y_1, y_2 \dots y_n)^t$	vector de la variable respuesta
$\vec{1}$	$=$	$(1, 1 \dots 1)^t$	vector de unos
$\vec{X}$	$=$	$(x_1, x_2 \dots x_n)^t$	vector de la variable regresora
$\vec{\varepsilon}$	$=$	$(\varepsilon_1, \varepsilon_2 \dots \varepsilon_n)^t$	vector de los errores aleatorios
$\hat{Y}$	$=$	$(\hat{y}_1, \hat{y}_2 \dots \hat{y}_n)^t$	vector de predicciones
$\vec{e}$	$=$	$(e_1, e_2 \dots e_n)^t = \vec{Y} - \hat{Y}$	vector de residuos

Dado el modelo de regresión

$$\vec{Y} = \alpha_0 \vec{1} + \alpha_1 \vec{X} + \vec{\varepsilon}$$

el método de estimación de mínimos cuadrados tiene la siguiente interpretación geométrica: el vector de predicciones $\hat{Y}$ es la proyección ortogonal del vector $\vec{Y}$ en el plano que generan los vectores $\vec{X}$ y $\vec{1}$. De esta forma el vector de residuos $\vec{\varepsilon}$ es mínimo. Y, por tanto, el vector de residuos es perpendicular al plano formado por $\vec{X}$ y $\vec{1}$, de donde:

$$\vec{e} \perp \vec{1} \;\; \Rightarrow \;\; \vec{e} \cdot \vec{1} = 0 \;\; \Rightarrow \;\; \sum_{i=1}^{n} e_i = 0$$

$$\vec{e} \perp \vec{X} \;\; \Rightarrow \;\; \vec{e} \cdot \vec{X} = 0 \;\; \Rightarrow \;\; \sum_{i=1}^{n} e_i x_i = 0$$

Tabla ANOVA: Contraste de regresión ó contraste conjunto de la F

En este apartado vamos a descomponer la variabilidad de la variable respuesta en variabilidad explicada por el modelo y variabilidad no explicada, lo que nos permitirá contrastar si el modelo es significativo o no. Esto es, bajo la hipótesis de que la relación que existe entre la variable respuesta y la regresora es lineal, estamos interesados en realizar el siguiente contraste de hipótesis:

$$H_0 : E(Y/X = 0) = \alpha_0 \quad \text{(el modelo no influye)}$$

frente a la alternativa

$$H_1 : E(Y/X = x) = \alpha_0 + \alpha_1 x \quad \text{(el modelo influye)}$$

Por tanto, si aceptamos H_0, la variable regresora no influye y no hay relación lineal entre las dos variables. En caso contrario, sí existe una dependencia lineal de la variable respuesta respecto a la regresora.

Para todos los datos muestrales podemos hacer la siguiente descomposición:

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$$

Se cumple:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

En base a esta igualdad se puede construir la tabla ANOVA mediante la cual resolver el **contraste de regresión de la F**.

Coeficiente de Determinación

En la interpretación del coeficiente de correlación debemos tener en cuenta que $R = \pm 1$ indica una relación lineal exacta positiva (creciente) o negativa (decreciente), $R = 0$ indica la no existencia de relación lineal estocástica, pero no indica independencia de las variables ya que puede existir una relación no lineal incluso exacta, y los valores intermedios indican la existencia de una relación lineal estocástica no exacta, más fuerte cuanto más próximo a ± 1 sea el valor de R.

Así pues, los pasos a seguir en Regresión Lineal son:

- Estimamos la mejor recta.
- Contrastamos Regresión y Linealidad.
- Medidos la bondad del ajuste ρ.
- Contrastamos las hipótesis de:
  - Linealidad.
  - Normalidad.
  - Homocedasticidad.
  - Independencia.
  - Outliers.
- Predecimos y/o estimamos.

Son causas de puntos atípicos:

- Punto observado con error en la medición, pero el modelo ajustado es adecuado.
- Punto observado es correcto pero el modelo ajustado no lo es por alguno de los siguientes motivos:
  - La relación entre las dos variables es lineal en un determinado intervalo pero donde se observa el punto no es lineal.
  - Hay una fuerte homocedasticidad que origina que algunas observaciones se separen de la nube muestral.
  - Existe una variable que no se tiene en cuenta en el modelo y que influye mucho en algunas observaciones.

Hipótesis de Independencia

Se estudia de la siguiente manera:

- √ Gráfico de residuos frente a tiempo (en orden de recogida).
- √ Cálculo de la función de autocorrelación muestral. Gráfico de autocorrelaciones.
- √ Contraste de Ljung-Box (χ^2 de Portmanteu).

igual que en diseño de experimentos. No obstante, en regresión lineal tenemos un contraste específico: el **contraste de Durbin-Watson**.

Contraste de Durbin-Watson

Es un constraste pensado para detectar errores que son dependientes con una estructura de autocorrelación AR(1), esto es, para el caso de que los errores sigan el siguiente modelo de dependencia:

$$\varepsilon_t = \rho\varepsilon_{t-1} + r_t$$

Las hipótesis del contraste son:

$$\begin{aligned} H_0 &\equiv \rho = 0 \quad \text{(independencia)} \\ H_1 &\equiv \rho \neq 0 \end{aligned}$$

El estadístico del contraste:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \simeq 2(1 - r_t)$$

siendo $e_t = y_t - \hat{y}_t$ los residuos y r_t la autocorrelación muestral de orden 1.

- Si $0 < d < d_L$, se rechaza H_0 y se acepta la existencia de autocorrelación positiva.
- Si $d_L < d < d_0$, el contraste no es significativo.
- Si $d_0 < d < 4 - d_0$, se acepta H_0 (no hay autocorrelación).
- Si $4 - d_0 < d < 4 - d_L$, el contraste no es concluyente.
- Si $4 - d_L < d < 4$, se rechaza H_0 y aceptamos la existencia de autocorrelación negativa.

Durbin y Watson calcularon la distribución de d bajo H_0 para cada n y cada α proporcionando d_L, d_0 niveles de significación superior e inferior de la distribución.

Modelo de Regresión Lineal General

Los nuevos problemas que se nos presentan son:

1. ¿Qué variables deben entrar en el modelo?
2. Una vez decididas las variables que entran en el modelo. . . ¿todas las variables introducidas en el modelo proporcionan nueva información? El uso/inclusión de variables que proporcionan la misma información que otras que ya están en el modelo da lugar a problemas de **multicolinealidad**; puede haber *mi-sespecificación* o existencia de *varios modelos válidos*.

Hipótesis del modelo:

En base a la var. de error ε_i	En base a la var. respuesta Y
E(ε_i)=0	E($y_i/x_{i1}x_{i2}\ldots x_{ik}$)=$\alpha_0+\alpha_1x_{i1}+\alpha_2x_{i2}\ldots$
Homocedasticidad Var(ε_i)=σ^2	Homocedasticidad Var($y_i/x_{i1}x_{i2}\ldots$)=σ^2
Independencia Cov($\varepsilon_i,\varepsilon_j$)=0 los errores ε_i son independientes	Independencia las observaciones y_i son independientes
Normalidad $\varepsilon_i \in N(0,\sigma)$	Normalidad $y_i/x_{i1}x_{i2}\ldots x_{ik}$ $\in N(\alpha_0+\alpha_1x_{i1}+\alpha_2x_{i2}\ldots+\alpha_kx_{ik},\sigma)$
$n>k+1$	$n>k+1$
las variables regresoras x_i son linealmente independientes	las variables regresoras x_i son linealmente independientes

La interpretación geométrica es análoga a la del Modelo de Regresión Lineal Simple:

$$\vec{1},\vec{x_1},\ldots,\vec{x_t} \Rightarrow \text{ subespacio vectorial } \mathbb{R}^{k+1}$$

por tanto

$$\hat{Y}-\alpha_0\vec{1}+\alpha_1\vec{x_1}+\ldots+\alpha_k\vec{x_k}$$

es la proyección ortogonal de $\vec{Y}$ en el subespacio vectorial engendrado, esto es,

$$\hat{Y}=V\vec{Y}$$

donde $\vec{V}$ es la *matriz de proyección* en el subespacio.

$$\left.\begin{array}{l}\vec{e}\perp\hat{Y}\\ \vec{e}\perp<\vec{1},\vec{x_1},\ldots\vec{x_k}>\end{array}\right. \Rightarrow \vec{e}\perp\vec{1},\vec{e}\perp\vec{x_1},\ldots\vec{e}\perp\vec{x_k}$$

de donde

$$\begin{array}{l}\sum e_i=0\\ \sum e_ix_{i1}=0\\ \vdots\\ \sum e_ix_{ik}=0\end{array}$$

$k+1$ condiciones

$n-(k+1)$ grados de libertad

Propiedades de los estimadores:

- El estimador mínimi-cuadrático $\hat{\alpha} = (X^t X)^{-1} X^t Y$ coincide con el estimador de máxima verosimilitud porque estamos bajo hipótesis de normalidad. ¿Cuál usar? Nos quedamos con el que tiene menor ECM.
- $\hat{\alpha}$ es insesgado (E($\hat{\alpha}$)=$\bar{\alpha}$).
- La varianza del estimador $\hat{\alpha}$ es

$$Var(\hat{\alpha}) = \sigma^2(X^t X^{-1}) = (\sigma^2_{ij})^k_{ij} = 0$$

- El estimador $\hat{\alpha}$ tiene distribución normal multivariante de orden k+1.
- El estimador $\hat{\alpha}_i$ tiene distribución normal:

$$\hat{\alpha}_i \in N(\alpha_i, \sigma\sqrt{g_i}) \qquad i = 0, 1, \ldots k$$

donde g_i es un término de la diagonal.

Pueden darse las siguientes situaciones:

Caso	Contraste conjunto de la F	Contraste individual de la t
1	Significativo	Todos significativos
2	Significativo	Algunos
3	Significativo	Ninguno
4	No significativo	Todos significativos
5	No significativo	Alguno
6	No significativo	Ninguno

Bibliografía

[1] Daniel Peña, Sánchez de Rivera. *Estadística. Modelos y métodos. Volumen 2: Modelos lineales y series temporales*. 2ª Edición Revisada, Alianza Universidad.

[2] Daniel Peña, Sánchez de Rivera. *Estadística. Modelos y métodos. Volumen 1: Fundamentos*. 2ª Edición, Alianza Universidad.