MODELOS LINEALES

Salvador Pintos

octubre/2002

Resumen

Notas para el curso de Métodos Estadísticos

Índice general

Capítulo 1

Modelos lineales

1.1. Introducción

El modelo lineal *análisis de regresión* es una metodología para examinar la asociación cuantitativa entre una variable de respuesta y con otras p variables de predicción x^j, donde se supone la existencia de n mediciones de la respuesta y_i, observadas bajo un conjunto de condiciones experimentales x_i^j de las variables de predicción.

Para cada i se formula una ecuación lineal:

$$y_i = \sum_{j=1}^{p} \beta_j x_i^j \quad + \varepsilon_i \qquad E(\varepsilon_i) = 0 \quad V(\varepsilon_i) = \sigma^2 \tag{1.1}$$

Donde los errores ε_i son variables aleatorias independientes.

El modelo es lineal en los parámetros β_j. Supuesta la validez del modelo, se estiman, por mínimos cuadrados, los parámetros β_j y la varianza σ^2 (que es un indicador del ajuste del modelo). Los estimadores $\widehat{\beta}_j$ de los parámetros β_j son óptimos en el sentido de ser insesgados y de varianza mínima. Una vez estimados los β_j, pueden hallarse los valores yp_i predichos por el modelo, asociados al conjunto de valores x_i^j de las variables de predicción, así como los residuales respectivos: $res_i = y_i - yp_i$.

Puesto que se obtienen estimadores $\widehat{\beta}_j$ de β_j, es necesario formular pruebas de hipótesis para determinar si los verdaderos valores β_j son nulos. Si un β_j es nulo, entonces el predictor x^j no debe pertenecer al modelo. Su exclusión es debida a que no está asociado con y, o, si lo está, su efecto ya está expresado

por las demás variables del modelo.

La elección de un buen modelo exige un análisis cuidadoso de los predictores a incluir en el modelo, así como de los residuales que deben ser independientes e igualmente distribuidos. Los parámetros estimados deben ser consistentes con el sentido físico del modelo.

1.2. Etapas en la construcción de un modelo

La metodología para la construcción de un buen modelo lineal se desarrolla en varias etapas:

1. Identificación de la estructura de entrada
2. Formulación y estimación del modelo
3. Análisis global
4. Análisis individual de los parámetros
5. Reformulación del modelo y volver al punto 2.
6. Seleccionar los mejores modelos alternativos
7. Estudio de residuales
8. Coherancia con la realidad
9. Elección del mejor modelo e interpretación del mismo
10. Predicción

1.2.1. Identificación de la estructura de entrada

Tormenta de ideas (reunión de expertos) para selección de posibles predictores.

Análisis de correlación entre las variables predictoras entre sí y con la respuesta. Análisis de componentes principales para identificar posibles sustituciones de variables porun conjunto reducido de componentes principales.

1.2.2. Formulación y estimación del modelo

Los aspectos matemáticos del Análisis de Regresión se sustentan en un manejo adecuado del álgebra matricial. Los paquetes estadísticos proporcionan los resultados de este análisis. Si se desea calcular personalmente los resultados fundamentales, debe disponerse de un programa donde el cálculo matricial sea amigable.

Vectores aleatorios

La noción de variable aleatoria se amplía a la de vector aleatorio:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

donde el valor esperado μ es el vector

$$\mu = \begin{bmatrix} E\,x_1 \\ E\,x_2 \\ \vdots \\ E\,x_p \end{bmatrix}$$

y la varianza de x está dada por la matriz cuadrada simétrica

$$V = E\left[(x-\mu)\,(x-\mu)^T\right]$$

que tiene en su diagonal la varianza individual de cada varaiable x_k y en cualquier elemento $V_{i,j} = cov(x_i, x_j)$.

Entre las propiedades elementales de vectores aleatorios aquí se usará:

- si c es un vector constante y x es aleatorio:

$$E(c^T x) = c^T\mu \quad Var(c^T x) = c^T V c \tag{1.2}$$

- si A es una matriz constante y x es aleatorio

$$E(Ax) = A\mu \quad Var(Ax) = AVA^T \tag{1.3}$$

- Si $F = x^T Ax$ es una forma cuadrática en x, entonces $E(F) = traza(AV) + \mu^T A\mu$

En lo que sigue se hará un sumario de los procedimientos matemáticos necesarios para la obtención de los resultados fundamentales del Análisis de Regresión, donde en cada fórmula los elementos con que se opera son matrices.

Resultados fundamentales.

Para usar el lenguaje sintético matricial es necesario expresar el conjunto de las n observaciones y_i como un vector columna y, del mismo modo los errores ε_i se agrupan en el vector columna ε y la matriz $(n \times p)$ de las observaciones x_i^j se representará como X. Es así que la ecuación 1.1 se reformula como:

$$y = X\beta + \varepsilon \quad con \quad E(\varepsilon) = 0 \quad V(\varepsilon) = \sigma^2 I \tag{1.4}$$

El vector ε tiene sus componentes estadísticamente independientes. Las dimensiones de y, X, β, y ε son: $n \times 1$, $n \times p$, $p \times 1$, y $n \times 1$ respectivamente. Las p $(p < n)$, variables independientes incluidas como columnas de X son determinísticas y constituyen un conjunto de vectores linealmente independientes, en cambio y es aleatoria ya que el vector ε lo es.

Si el modelo incluye una constante (por ejemplo, si el modelo es: $Y = a + bx_1 + cx_2$), la matriz X debe tener 3 columnas, donde su primer columna está formada por unos y las siguientes por los valores de x_1 y x_2.

La figura 1.1 presenta dos modelos alternativos para una data de la población de EEUU en función de los años: uno lineal $Pob = \beta_0 + \beta_1 año + \varepsilon$ cuyos residuales se analizarán más adelante y el otro cuadrático $Pob = \beta_0 + \beta_1 año + \beta_2 año^2 + \varepsilon$. Como es obvio la apreciación visual no es suficiente para optar entre modelos alternativos, y si lo fuera quedaría limitada a modelos extremadamente simples, con a lo sumo dos variables de entrada.

Nótese además, que el modelo es lineal debido a que es *lineal en los parámetros*, a pesar de que una de las variables de entrada sea el cuadrado de otra.

1.2.3. Estimación de parámetros

Los parámetros involucrados en el modelo β y σ^2 deben ser estimados. Así mismo se desea poder predecir con el modelo identificado.

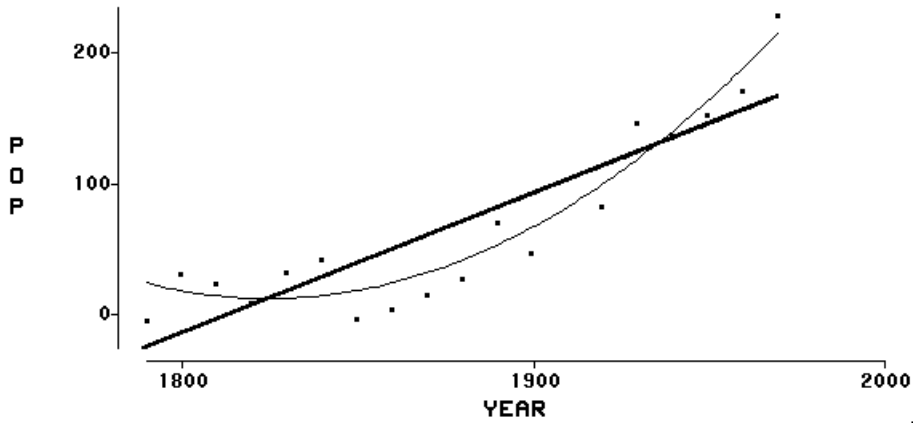

Figura 1.1: Modelo parabòlico

La resolución de la ecuación 1.4 por el método de mínimos cuadrados consiste en minimizar en β la función:

$$SSE = \min_{\beta} \|y - X\beta\|^2 \tag{1.5}$$

Derivando la ecuación 1.5 respecto de β resulta que: $X^T X\beta = X^T y$, y si las columnas de la matriz X son independientes, $X^T X$ es invertible y la solución está dada por:

$$\widehat{\beta} = \left(X^T X\right)^{-1} X^T y \tag{1.6}$$

el estadístico $\widehat{\beta}$ tiene la propiedad de ser el estimador insesgado y de varianza mínima de β. Aplicando 1.3pruebe que $E\left(\widehat{\beta}\right) = \beta$.

Sea yp el vector con las predicciones de las n observaciones obtenidas con el modelo:

$$yp = X\widehat{\beta} = X\left(X^T X\right)^{-1} X^T y = Py \tag{1.7}$$

donde $P = X\left(X^T X\right)^{-1} X^T$ es la matriz de proyección de un vector cualquiera sobre el espacio de las columnas de X. Por ser P una matriz de proyección es idempotente $(P = P * P)$ y simétrica $P = P^T$. Para toda observación nueva, $xnew$, su predicción es simplemente $xnew^T\widehat{\beta}$.

El mínimo de la suma de cuadrados de los errores es

$$SSE = \left\|y - X\widehat{\beta}\right\|^2 = \|y - yp\|^2 = y^T\left(I_n - P\right)y \tag{1.8}$$

donde I_n es la matriz identidad de tamaño n. El estimador de la varianza σ^2 está dado por el promedio de los errores, el error medio cuadrático MSE:

$$MSE = \widehat{\sigma^2} = \frac{SSE}{n-p} \tag{1.9}$$

donde se divide por $n-p$ para obtener una estimación insesgada de σ^2.

Precisión de los estimadores

Aplcando 1.3 se deduce la matriz de covarianza del estimador $\widehat{\beta}$:

$$COV\left(\widehat{\beta}\right) = \left(X^T X\right)^{-1} \sigma^2 \tag{1.10}$$

como no se conoce σ^2 en el cálculo anterior se la sustituye por $\widehat{\sigma^2}$. La varianza de cada parámetro individual se encuentra en la diagonal de $COV\left(\widehat{\beta}\right)$.

Igualmente de 1.3 se deduce que:

$$COV(yp) = P\sigma^2 \tag{1.11}$$

1.2.4. Ajuste global del modelo

El modelo lineal suele contener una constante además de las variables predictores, es decir:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j \quad + \varepsilon$$

Para aplicar los resultados ya expuestos se incluye en la matriz de diseño, X, una columna de unos que multiplican a β_0. Si el modelo sólo tuviera esta variable su estimador $\widehat{\beta_0} = \overline{y}$. Es por ello que para aislar este estimador de las variables predictoras la suma de cuadrados total $\|y\|^2$, se modifica y se la calcula respecto de la media $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ de la respuesta. La suma de cuadrados total corregida de las observaciones respecto de la media $\overline{y}$, está dada por $SST = \|y - \overline{y}\|^2 = \|y\|^2 - n\overline{y}^2$. Esta suma se descompone en la suma de cuadrados que expresa el modelo: SSM más la suma de cuadrados de los residuales SSE.

$$SST = SSM + SSE \tag{1.12}$$

La descomposición en suma de cuadrados permite analizar el efecto global del modelo; como SST no depende del modelo, para que el modelo ajusta bien

debe reducirse SSE o incrementarse SSM. Los indicadores del ajuste global del modelo son:

- El error medio cuadrático MSE
- El coeficiente de deteminación *(R-square)* R^2
- El coeficiente de deteminación ajustado $AdjR^2$

El error medio cuadrático (*Mean Square Error*) es el estimador insesgado de la varianza σ^2 del modelo. Su raíz cuadrada $RMSE$, es el estimador de la desviación estándar σ del modelo. De la ecuación 1.8 se obtiene:

$$MSE = \frac{SSE}{n-p} = \frac{y^T (I_n - P)\, y}{n-p} \quad RMSE = \sqrt{MSE} \tag{1.13}$$

Cuanto más pequeños sean estos estimadores mejor es el modelo.

El coeficiente de deteminación R^2 explica la relación entre la suma de cuadrados que expresa el modelo y la suma de cuadrados total. De la ecuación 1.12 se obtiene:

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST} \tag{1.14}$$

Se cumple que $0 \leq R^2 \leq 1$ y el ajuste del modelo se observa por la proximidad de R^2 a 1.

Este indicador tiene el defecto de no penalizar el exceso de variables, ya que cada vez que se agrega una nueva variable al modelo, el R^2 crece sin que esto signifique que la nueva variable aporte algo al modelo. Para evitar el exceso de parámetros se introduce un nuevo indicador.

El *coeficiente de determinación ajustado,* $AdjR^2$, penaliza el aumento del número p de variables y se deriva de la ecuación 1.14.

$$AdjR^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = R^2 - \frac{p-1}{n-p}\left(1 - R^2\right) \tag{1.15}$$

El $AdjR^2$ resta de 1 el cociente entre dos varianzas estimadas: la del modelo, MSE y la de las observaciones, y_i. Cuando el modelo ajusta bien, el cociente debe ser próximo a 0 y el $AdjR^2$ proximo a 1.

Este último coeficiente proporciona una información más adecuada que el R^2 ya que penaliza el exceso de predictores en el modelo y uno de los propósitos de modelar es que el modelo sea parsimonioso (cuanto más simple mejor).

1.2.5. Análisis individual de los parámetros

Para hacer el estudio individual de cada uno de los parámetros del modelo es necesario conocer qué propiedades satisfacen. Si el error ε del modelo en la ecuación 1.4 se distribuye *normalmente*, entonces, los estadísticos $\widehat{\beta}$ y yp también son normales y es posible hacer pruebas de hipótesis acerca de sus valores.

Para que una variable x_i pertenezca al modelo es necesario que su coeficiente asociado, β_i, sea distinto de 0. Como se ha estimado el vector β mediante $\widehat{\beta}$, es necesario realizar una prueba de hipótesis de los parámetros del modelo, con un cierto nivel de riesgo (15 %, por ejemplo), para saber realmente si $\widehat{\beta}_i \neq 0$.

La desviación estándar de los elementos de $\widehat{\beta}$, se obtiene de las ecuaciones 1.10 y 1.13, y está dada por: $\widehat{\sigma_\beta} = RMSE\sqrt{diag\,(X^TX)^{-1}}$, donde *diag* significa la diagonal de una matriz. Para cada variable x_i el cociente

$$tratio_i = \frac{\widehat{\beta}_i}{\widehat{\sigma_\beta}_i} \tag{1.16}$$

se comporta como una distribución t -*Student*- con $n - p$ grados de libertad. Si el $tratio_i$ supera el valor t asociado al nivel de significación entonces el verdadero valor β_i es distinto de 0, y la variable asociada debe pertenecer al modelo. De otra manera si el valor-p asociado a el $tratio_i$ es menor que el nivel de significación α del test la variable debe pertenecer al modelo.

Los errores de primer tipo α para un modelo de regresión se ubican habitualmente en un valor del 15 %, Si se es muy exigente (por ejemplo $\alpha = 0{,}05$), se corre el riesgo de que no queden variables en el modelo. Si se tiene muchas variables en el modelo y se desea reducir su número entonces se disminuye α.

1.2.6. Reformulación del modelo

Si en las pruebas de hipótesis alguna variable es no signicativa es necesario excluirla y realizar las etapas 2, 3, y 4 nuevamente. Las variables en un modelo de regresión deben ser eliminadas de a una, ya que la presencia de dos variables altamente correlacionadas puede hacer que éstas "se estorben" y ambas figuren para ser eliminadas, sin embargo, cada una de ellas incluidas únicamente pueden ser significativas. Entre las no significativas se elige la que tiene el $tratio$ más cercano a 0 . Luego, se recalcula totalmente el modelo.

Para cada modelo es necesario guardar sus indicadores de ajuste global, ya que entre todos los modelos es necesario seleccionar aquellos de mejor ajuste.

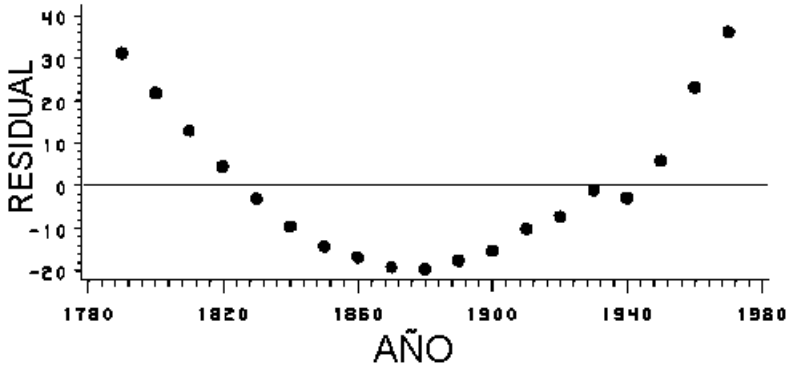

Figura 1.2: Residual respecto del modelo lineal

1.2.7. Seleccionar los mejores modelos alternativos

Luego de repetir varias veces el proceso anteriormente descrito se seleccionan los mejores modelo, (2 o 3) para un análisis más profundo, mediante la comparación de los estadísticos de ajuste global. Algunos productos estadísticos tienen programas que realizan automaticamente el estudio de selección, pero esta selección se lleva a cabo optimizando unilateralmente alguno de los criterios expuestos, (por ejemplo el R^2 ajustado) para seleccionar un buen modelo.

1.2.8. Estudio de residuales

El residual, $res = y - yp$ es la diferencia entre los valores los valores observados y las respectivas predicciones yp .

Se grafican los residuales contra cada una de las variables de entrada ($res\ vs\ x_i$) en busca de patrones. Si el residual cumple con la hipótesis de ruido blanco habrá ausencia de patrones en estos gráficos. La figura 1.2es el residual de un modelo muy simple de la población de EEUU: $Pob = \beta_0 + \beta_1 año + \varepsilon$, el residual muestra un comportamiento cuadrático. En ese caso se sugiere agregar un término adicional con el cuadrado del año:

$Pob = \beta_0 + \beta_1 año + \beta_2 año^2 + \varepsilon$. Si los residuales no tienen un comportamiento adecuado es posible que sea necesario incluir nuevas variables anteriormente no previstas.

1.2.9. Coherancia con la realidad

Los parámetros finalmente estimados en el modelo óptimo deben ser coherentes con el comportamiento esperado del modelo.Por ejemplo, no es posible que un modelo de ventas de un producto en función del precio, tenga el coeficiente estimado para la variable precio positivo. En ese caso el modelo debe ser desechado y volver a alguno de los modelos subóptimos y analizar su coherencia.

1.2.10. Elección del mejor modelo e interpretación del mismo

Una vez que se ha optado por un modelo éste debe interpretarse. Para realizar un estudio de sensibilidad se analiza la influencia de cada variable de entrada en la respuesta. La derivada parcial de la variable respuesta respecto de una variable independiente representa el incremento de la respuesta por unidad de cambio de la variable de entrada. Si la variable independiente está "sola", esta derivada es el propio paramétro estimado.

1.2.11. Predicción

Los modelos de regresión tiene la virtud de establecer intervalos de confianza de la predicción estimada. La desviación estándar de una predicción, yp,tomada de la data original está dada por (ver ecuación 1.11):

$$std(yp) = RMSE\sqrt{diag(X\left(X^TX\right)^{-1}X^T) + 1} \tag{1.17}$$

El 1 agregado al cálculo de la varianza proviene de la varianza del residual ya que lo que observamos es $y = yp + res$.

Luego, dada la confianza $1-\alpha$, sea $tsup$ el valor que sobre la distribución t de $n-p$ grados de libertad deja en el extremo derecho de la curva de densidad un área $\alpha/2$. Un intervalo de $1-\alpha$ de confianza de la predicción está dado por :

$$[yp - tsup\ std(yp),\ yp + tsup\ std(yp)] \tag{1.18}$$

La figura 1.3 presenta una producción ficticia hasta el año 1970, la predicción para un horizonte de 50 años y los intervalos de confianza respectivos. Obsérvese que los intervalos de confianza se abren en la medida que se predice alejándose del presente.

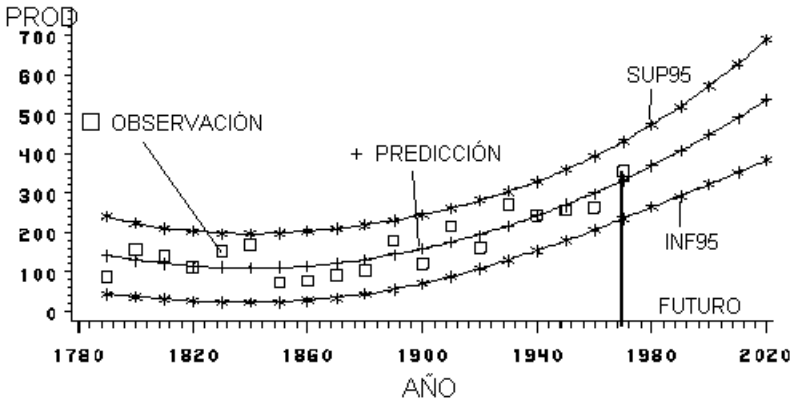

Figura 1.3: Intervalos de confianza de la predicciòn

Cuando se considera una nueva observación z y se desea predecir la respuesta del modelo para este valor se tiene:

$$Y_z = z^T \widehat{\beta} = z^T \left(X^T X\right)^{-1} X^T y$$

y la varianza de esta estimación es la suma de la varianza de $z^T \widehat{\beta}$ más la varianza del residual del modelo:

$$V(Y_z) = \widehat{\sigma^2} \left(z^T \left(X^T X\right)^{-1} z + 1\right) = \frac{SSE}{n-p} \left(z^T \left(X^T X\right)^{-1} z + 1\right)$$

1.3. Test de hipótesis

Si se desea probar el test $L\beta = 0$,donde L es una matriz de dimensionessxp de rango $s < p$, el estadístico:

$$f = \frac{(Lb)^T \, inv\left(L \left(X^T X\right)^{-1} L^T\right) Lb}{s \, MSE}$$

se comporta como una F de s y $n-p$ grados de libertad. Luego, el valor p del test esta dado por

$$vp = 1 - probF(f, s, n-p)$$