APLICACIONES DE LOS MODELOS DE PRONÓSTICOS EN SERIES DE TIEMPO

Noé Panozo Jiménez
Carrera de Ingeniería de Sistemas, Escuela Militar de Ingeniera
La Paz, Bolivia
npanozoj@adm.emi.edu.bo

APPLICATIONS OF PROJECTS MODELS IN TIME SERIES

Resumen— El presente trabajo tiene por objetivo proponer un uso adecuado de los modelos de Pronósticos en Series de Tiempo, con comportamientos estocásticos.

Abstract— The objective of this paper is to propose an adequate use of the Forecast Models in Time Series, with stochastic behaviors.

I. INTRODUCCION

En muchas de sus aplicaciones, la dependencia entre las observaciones se mira como un estorbo y en la planeación o diseño de experimentos, la aleatorización se trabaja para validar el análisis como si las observaciones fueran independientes. Lo más común es que la forma funcional f de la serie y sus coeficientes sean desconocidos y, por lo tanto, deban ser estimados a partir de datos históricos. El análisis de series de tiempo sólo usa el historial secuencial y ordenado de los datos observados de la serie de la variable que se va a pronosticar, para desarrollar un modelo para predecir valores futuros. Si estas otras variables están correlacionadas con la variable de interés y si parece que hubiera alguna causa de esta correlación, se puede construir un modelo estadístico que la describa.

Lo más común es que la forma funcional f de la serie y sus coeficientes sean desconocidos y, por lo tanto, deban ser estimados a partir de datos históricos.

Su uso comprueba que la posibilidad de componentes periódicos de frecuencia desconocida, pueden permanecer en la serie. Pueden ser usados para derivar políticas de control óptimas que muestren cómo puede manejarse una variable bajo control, para minimizar perturbaciones en algunas variables dependientes.

La obtención de buenos pronósticos se ha convertido en elemento clave a través de la historia.

Según la disponibilidad de los datos, una relación de pronósticos puede suponerse como una función del tiempo, o de variables independientes, y comprobarse. Un paso importante en la selección del método apropiado es el de considerar aquellos patrones que pueden ser comprobados. Se pueden distinguir cuatro tipos de patrones de datos:

1. Un patrón horizontal (H) existe cuando los datos fluctúan alrededor de una media constante. (tal serie es estacionaria en su media). Por ejemplo, un proceso cuya situación de control de calidad no cambie, es de esta clase.

2. Un patrón estacional (S) existe cuando una serie se ve influenciada por factores estacionales.
3. Un patrón cíclico (C) se presenta cuando los datos se ven afectados por fluctuaciones en el tiempo.
4. Existe un patrón de tendencia (T) cuando hay un aumento o decrecimiento secular a largo plazo en los datos.

Muchas series de datos incluyen combinaciones de los patrones anteriores. Si se necesita una desagregación de los componentes de los patrones, los métodos de pronóstico pueden ser útiles para distinguir cada uno de ellos. Análogamente, estos métodos pueden usarse para identificar el patrón y el mejor ajuste de los datos para pronosticar el valor futuro.

II. MÉTODOS DE PRONÓSTICO

El objetivo de los pronósticos es el de reducir el riesgo en la toma de decisiones. Dedicándole más recursos, se debe mejorar la exactitud, y en consecuencia eliminar algunas de las pérdidas resultantes de la incertidumbre en la toma de decisiones. Debido a que nunca se puede eliminar el riesgo de los pronósticos, es necesario que el proceso de decisión considere explícitamente la incertidumbre latente en el pronóstico. A menudo, la decisión se relaciona conceptualmente con el pronóstico por DECISION REAL = DECISION BAJO LA HIPOTESIS DE QUE EL PRONOSTICO ES CORRECTO + TOLERANCIA DEL ERROR DEL PRONOSTICO.

El método de pronósticos es una parte de un sistema complejo de administración y como subsistema, interactúa con otros componentes del sistema global para determinar su desempeño. Es difícil medir la utilidad de los pronósticos. Son usados, principalmente, para hacer insinuaciones, ayudar a planear y complementar pronósticos cuantitativos mejor aún que suministrar un dato numérico. Debido a su naturaleza y costo, se usan casi exclusivamente para situaciones a medio y largo plazo para formular estrategias, desarrollo de nuevos productos y tecnologías y planes a largo plazo.

Para determinar un pronóstico se debe empezar definiendo las variables que deben analizarse y pronosticarse. El nivel de lo detallado del proceso es algo muy digno de tenerse en cuenta. Un sistema de planeación de producción puede requerir un pronóstico de la demanda en unidades para cada producto terminado para programar el uso de las instalaciones y el nivel de inventarios. Por otro lado, un gerente de ventas puede requerir sólo un pronóstico de las ventas totales en pesos como información para su presupuesto.

En la planeación de producción se puede proceder en forma de valores agregados, como sería el caso de familias de productos similares para luego desglosar los pronósticos calculados a nivel de unidades en una segunda instancia. De ahí lo importante del dominio que se debe tener sobre el manejo de números índices

El objetivo de los pronósticos es el de reducir el riesgo en la toma de decisiones.

En algunos casos el término tiempo muerto de pronóstico se usa en lugar de horizonte del pronóstico. Puesto que los pronósticos resultan menos precisos al aumentar el horizonte, se puede, a menudo, mejorar el proceso decisorio acortando el tiempo muerto, en consecuencia, reduciendo el tiempo muerto del pronóstico y permitiendo una reacción más rápida al error del pronóstico. Otro aspecto de este tema se refiere a la forma requerida del pronóstico. Siempre es conveniente pensar en la variable de interés como una variable aleatoria con una distribución desconocida de probabilidad.
En ciertos casos puede haber más interés en predecir cambios significativos en el proceso que da origen a una variable que en predecir el valor de esa variable. Los datos históricos son muy valiosos para establecer los procedimientos de pronósticos y las observaciones futuras deben estar presentes para la revisión de los pronósticos. Además, se debe examinar la representatividad de los datos. Entonces, la empresa debe implantar un procedimiento de recolección de datos que indique lo que los

clientes desean realmente que se les envíe en cada período de entrega.

Esta es la clase de pronóstico requerido para planear diferentes líneas de producto. Se puede esperar que los datos estén correlacionados con los presupuestos en muchas empresas. La presencia de un juicio en el proceso de pronósticos es importante, pero requiere personal interesado en participar.

Otro factor importante es el de la disponibilidad de los datos. Los datos históricos son muy valiosos para establecer los procedimientos de pronósticos y las observaciones futuras deben estar presentes para la revisión de los pronósticos.

El objetivo es obtener una función de pronósticos tal que la media cuadrática de las desviaciones $z_{t-1} - z_1$, sea tan pequeño como sea posible para cada tiempo muerto. Fuera de calcular los mejores pronósticos, es necesario especificar su precisión de tal manera que se puedan calcular los riesgos asociados con las decisiones basadas en los pronósticos ya calculados. Esto se puede lograr obteniendo los límites probabilísticos a ambos lados de cada pronóstico. Los métodos para hacer pronósticos se clasifican ampliamente en cualitativos y cuantitativos.

Generalmente estos pronósticos no pueden ser reproducidos por alguien más, puesto que el pronosticador no indica claramente cómo fue incorporada la información disponible en el pronóstico. Aunque et tema de los pronósticos subjetivos no es un enfoque riguroso, puede ser muy apropiado y el único método razonable en muchas situaciones. Una vez se seleccione el modelo o la técnica, se pueden determinar automáticamente los pronósticos correspondientes para ser reproducidos cuando se necesiten.

$$Y = f(X_1, X_2, \ldots, X_p) + \text{ruido (o perturbación)} \quad [1]$$

Las observaciones estadísticas usadas para determinar el modelo, es lo que se conoce como datos empíricos y son el tema principal de estudio en este artículo. Por ejemplo, se puede hacer un estudio del comportamiento de una serie sobre el aumento del costo de la canasta familiar y extrapolar el modelo hacia los próximos meses y luego verificar la veracidad del modelo como tal. Para lograr este objetivo se pueden usar métodos de suavización o modelos paramétricos de series cronológicas. Para hacer pronósticos a través de la regresión, se hace uso de la relación entre la variable a pronosticar y las otras variables que explican su variación.

Es decir, se pueden hacer pronósticos mensuales de las ventas de textiles a partir de su precio, el ingreso disponible de los consumidores y los gastos en publicidad. En los demás casos se debe tener en cuenta la influencia de otras variables relacionadas con la de interés, para ver el comportamiento futuro mediante un modelo de regresión dinámica o función de transferencia. Para construir un modelo se debe disponer de datos relevantes y una teoría científica de soporte. Es aquí donde los datos juegan un papel predominante para determinar la clase de modelo.

Luego, el analista debe verificar lo adecuado del modelo ajustado. Si el modelo es insatisfactorio, necesita ser replanteado y el ciclo iterativo del modelo especificación – estimación - diagnóstico - verificación debe ser repetido hasta encontrar un modelo satisfactorio. Puesto que los pronósticos dependen absolutamente del modelo, debe haber seguridad de que éste y sus parámetros permanezcan constantes durante el período de su aplicación como tal. At calcular las desviaciones, o errores del pronóstico, se detecta cualquier cambio en el modelo.
Es decir, las funciones particulares de estos errores pueden indicar un sesgamiento en los pronósticos indicando unas predicciones sub o sobre valoradas. Además, se puede hacer uso de las observaciones más recientes para actualizar los pronósticos. Debido a que las observaciones se llevan secuencialmente en el tiempo, los procedimientos de actualización se pueden aplicar de manera rutinaria evitando los cálculos de cada pronóstico desde el principio.

III. SELECCIÓN DE UN MODELO PARTICULAR DE PRONÓSTICOS

Aunque en algunas ocasiones sólo se necesita obtener pronósticos en bruto, en otros casos se requiere gran precisión. El analista debe construir modelos simples porque son más fáciles de entender, usar y explicar. Un modelo muy pulido puede conducir a predicciones más precisas, pero puede ser más costoso y difícil de implementar.

IV. CRITERIOS DE PREDICCIÓN

Ahora, el objetivo es encontrar un pronóstico tal que el error futuro $Z_t - Z_{t-1}$ sea tan pequeño como sea posible. Entonces, se puede hablar sólo de su valor esperado condicionado a los datos históricos hasta el instante t-1 inclusive. Si ambos errores de pronóstico, el negativo y el positivo son igualmente indeseables, se justifica elegir el pronóstico de tal manera que la media absoluta del error $E[Z_t - Z_{t-1}]$ o el error medio cuadrático $E[Z_t - Z_{t-1}]^2$ sean mínimos. Los pronósticos que minimizan este valor, se conocen con el nombre de pronósticos del error medio cuadrático mínimo o EMCM.

V. EL MODELO DE REGRESIÓN Y SU APLICACIÓN EN PRONÓSTICOS

El análisis de regresión tiene como objetivo principal encontrar el modelo apropiado para describir la relación entre dos o más variables y, además, cuantifica el grado de respuesta de la variable dependiente ante diversos cambios de las variables predictoras.

Pero como la realidad es bien diferente, se debe confiar en observaciones empíricas para obtener resultados aproximados. El alcance de los modelos es esencial puesto que los métodos que producen pronósticos a corto plazo difieren de los de largo plazo.

VI. EL MODELO DE REGRESIÓN

Muchos de los modelos usados para representar series cronológicas son algebraicos o funciones trascendentales (los modelos que contengan términos trigonométricos o exponenciales, se llaman a menudo modelos trascendentales) del tiempo o algún modelo compuesto que combine ambos componentes. Por ejemplo, si las observaciones son muestras aleatorias de alguna distribución de probabilidad, y si la media de la distribución no cambia con el tiempo, se puede usar el modelo constante:

$$x_t = b + E_t \qquad [2]$$

donde x_t es la demanda en el período t, b es la media desconocida del proceso y E_t es el componente aleatorio. Si se supone que la media del proceso cambia linealmente con el tiempo, se usa el modelo de tendencia lineal:

$$x_t = b_1 + b_{2t}E_t \qquad [3]$$

También se puede expresar una variación cíclica introduciendo términos trascendentales en el modelo; por ejemplo:

$$x_t = b_1 + b_2 sen\frac{2\pi t}{12} + b_3 cos\frac{2\pi t}{12} + E_t \quad [4]$$

la cual se refiere a un ciclo que se repite cada 12 períodos.

La forma general de estos modelos es:

$$x_1 = b_1\, z_1(t) + b_2\, z_2(t) + \cdots + b_k\, z_k + E_t \qquad [5]$$

donde los $\{b_t\}$ son los parámetros, las $\{z_t(t)\}$ son funciones matemáticas de t y E_t es el componente aleatorio. Nótese que este enfoque de modelos representa el valor esperado del proceso como una función matemática de t. A menudo es deseable definir el origen del tiempo como el final del período más reciente T. Entonces el modelo para la observación en el período $T = T + \tau$ lo indica [6]:

$$x_{t+\tau} = a_1\,(T)z_1(\tau) + a_2\,(T)z_2(\tau) + \cdots + a_k\,(T)z_k(\tau) + E_{t+\tau} \qquad [6]$$

en la cual se denotan los coeficientes por $\{a_t(T)\}$ para indicar que se basan en el origen actual en el tiempo y que se distinguen de los coeficientes del origen ordinario. El hecho de mantener el origen del tiempo en una base

actualizada facilita enormemente el trabajo en un sistema de pronósticos.

Otro modelo muy diferente es aquél cuyas observaciones x_t actúan como una función de componentes aleatorios previos E_t, E_{t-1}, E_{t-2}, ….., en la forma general [7]:

$$x_t = \mu + \theta_0 E_t + \theta_1 E_{t-1} + \theta_2 E_{t-2} + \cdots \quad [7]$$

donde μ y $\{\theta_t\}$ son constantes. Los modelos de esta clase reciben comúnmente el nombre de modelos de filtro lineal. Puede que no sea obvio que un modelo de esta forma pueda representar una serie de tiempo, aunque se verá que bajo este enfoque se pueden obtener resultados excelentes. Esto es particularmente válido para series cuyas observaciones son altamente correlacionadas, es decir, cuando las observaciones no son mutuamente independientes.

Una técnica sencilla para seleccionar un modelo para aplicar una serie de tiempo, consiste en graficar los datos históricos en un diagrama de dispersión e identificar la función matemática que más se les parezca. Puesto que el modelo debe representar el futuro cercano con miras a pronosticar, generalmente se juzga su efectividad identificando la descripción de su pasado más reciente. La simulación es una técnica muy útil para evaluar los métodos alternativos de pronósticos. Esto se puede hacer retrospectivamente usando datos históricos.

Si parece que el futuro fuese a diferir del pasado, se puede crear una pseudo historia de las expectativas subjetivas de la naturaleza futura de la serie y usarla en la simulación. La simulación también es útil para determinar los parámetros de las técnicas de pronósticos, y para obtener mejores constantes de suavización. Es conveniente pensar en dos funciones primarias de un sistema de pronósticos como son la generación y el control de pronósticos. El control de los pronósticos involucra el monitoreo del proceso para detectar condiciones fuera de control e identificar las oportunidades para mejorar el desempeño del modelo.

Este estadístico se calcula dividiendo un estimador del error estimado del pronóstico por una medida de la variabilidad del error del pronóstico, tal como el estimador de la desviación absoluta media del error del pronóstico. Una función de control de los pronósticos también debe involucrar informes periódicos sobre su utilidad y presentar los resultados a la administración respectiva. Esta retroalimentación debe motivar un mejoramiento en los aspectos cuantitativos y cualitativos del sistema.

Una técnica sencilla para seleccionar un modelo para aplicar una serie de tiempo, consiste en graficar los datos históricos en un diagrama de dispersión e identificar la función matemática que más se les parezca.

Sea la relación entre una variable dependiente Y_t y p variables independientes X_1, X_2, X_3,…, X_p ; si se agrega un índice t, el valor de la variable dependiente en el instante t (o para el item t) se designará por y_t y las p variables independientes tendrán los símbolos x_{t1}, x_{t2},…., x_{tp}. Si las observaciones ocurren secuencialmente en el tiempo, el índice t se refiere a tiempo; en cualquier otro caso, t es simplemente un índice arbitrario. Es decir, en tablas de doble entrada, por ejemplo, donde las observaciones se refieran a elementos diferentes (empresas, regiones, etc.), el orden no tiene significado alguno. En su forma general, el modelo de regresión se expresa:

$$y_t = f(x_t; \beta) + E_t \quad [8]$$

donde f (x_t; β) es una función matemática de las p variables independientes x_t = (x_{t1},……,x_{tp}) y de los parámetros desconocidos β= (β_1,…., β_m). En el estudio del modelo se supone que la forma funcional es conocida, aunque los parámetros sean desconocidos. El modelo es probabilístico, puesto que el término de error es una variable aleatoria y se supone que sigue una distribución normal, que sus valores son no correlacionados (es decir, independientes entre sí) o sea que Cov (E_t,…..,E_{t-k}) = E(E_t,…..,E_{t-k}) = 0, para todo t y k ≠ 0, y su esperanza es cero y varianza σ^2. En

notación vectorial, se tiene que el vector columna de los errores es igual a cero y la matriz de covarianzas es simétrica de tamaño n x n e igual σ^2 x 1.

Debido a que el modelo es aleatorio, se concluye que la variable dependiente, y_t, también lo es. Por lo tanto el modelo puede ser expresado en forma de distribución condicional de y_t dado $x_t = (x_{t1},\ldots, x_{tp})$ Bajo este contexto, las hipótesis pueden expresarse como:

a. La media condicional $E(y_t/x_t) = f(x_t; \beta)$ depende de las variables independientes x_t y de los parámetros β, y la varianza $V(y_t/x_t) = \sigma^2$ es independiente de x_t y del tiempo.

b. Las variables dependientes y_t y y_{t-k} para diferentes instantes o temas, son incorrelacionados:

$$Cov(y_t, y_{t-k}) = E[y_t - f(x_t;\beta)][y_{t-k} - f(x_{t-k};\beta)]=0 \quad [9]$$

c. El panel condicional de y_t sobre x_t sigue una distribución normal con media $f(x_t; \beta)$ y varianza σ^2 la cual se denota por $N[f(x_t; \beta), \sigma^2]$.

Estas hipótesis implican que la media de la distribución condicional de y_t es una función de la variable independiente x_t. Sin embargo, esta relación no es determinística porque para valor fijo de x_t habrá una dispersión de los y_t alrededor de su media; aún más, el error no puede ser expresado con anticipación.

Debido al enfoque resumido de este artículo y a lo accesible del tema en la bibliografía, dejan de presentarse aquí los diferentes modelos y sus funciones, la deducción de las fórmulas para estimar los parámetros mediante el uso de las ecuaciones normales, o sea la aplicación del método de los mínimos cuadrados en su forma matricial, y las propiedades de los estimadores mínimos cuadráticos.

VII. PREDICCIÓN USANDO LOS MODELOS DE REGRESIÓN CON COEFICIENTES ESTIMADOS

Cuando se desconocen los parámetros $\beta = (\beta_0, \beta_1,\ldots, \beta_k)$ se reemplazan por sus estimadores mínimos cuadráticos, expresados aquí en forma matricial: $\beta = (XX)^{-1}$ X'Y. Así, el valor de predicción de y_t está dado por:

$$\hat{y_k^{pred}} = \hat{\beta_0} + \hat{\beta_1} x_{k1} + \cdots + \hat{\beta_p} x_{kp} = x_k' \hat{\beta} \quad [10]$$

Se puede demostrar fácilmente que estas predicciones tienen las siguientes propiedades:

a. El pronóstico obtenido de esta manera es un predictor insesgado puesto que el valor esperado de un error de pronóstico futuro es cero.

b. Aún más, $y_k^{\hat{}pred}$ es el pronóstico de error mínirno cuadrático entre todos los pronósticos lineales insesgados. Esto se puede comprobar a partir del teorema de Gauss-Markov.

c. La varianza del pronóstico o equivalentemente del pronóstico está dada por [11]:

$$V(y_k - \hat{y_k^{pred}}) = V[E_k + x_k'(\beta - \hat{\beta})]\hat{y_k^{pred}} = \sigma^2[1 + x_k'(X'X)^{-1}x_k] \quad [11]$$

Aquí se ha usado $V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$ y se ha supuesto que el error futuro E_k no está correlacionado con los errores en el período de estimación.

d. La varianza estimada del error del pronóstico

$$\hat{V}(y_k - \hat{y_k^{pred}}) = S^2[1 + x_k'(X'X)^{-1}x_k] \quad [12]$$

donde $S^2 = SSE/(n - p - 1)$ es el error medio cuadrático que sirve como factor para la estimación de los límites del intervalo de predicción de un valor futuro y_k al 100 (1- α)% de confianza.

VIII. TÉCNICAS PARA LA SELECCIÓN DE MODELOS

La selección de las variables independientes que deben ser incluidas en el modelo definitivo es

uno de los problemas más importantes de resolver en la construcción empírica de modelos.

Generalmente, al principio del proceso se puede disponer de una lista muy extensa de las posibles variables "predictoras", pero nadie está seguro de cuántas y, especialmente cuáles, deben incluirse en el modelo definitivo. Obviamente, no es del caso omitir variables importantes explicativas. Pero, por otro lado, se pretende mantener el modelo lo más simplificado posible para que sea más fácil de entender y de explicar. Además, se debe tener en cuenta que la estimación de parámetros innecesarios introduce incertidumbre adicional a los pronósticos.

Un enfoque aceptable consiste en estimar todos los modelos de regresión posibles. Sin embargo, para k variables, se requiere estimar 2^k modelos.

El error medio cuadrático s^2, el coeficiente de determinación R^2 , o preferiblemente su ajustado, R_a^2, donde:

$$R_a^2 = 1 - [\frac{SSE}{n-p-1}]/[\frac{SSTO}{n_1}] \qquad [13]$$

puede ser calculado para cada subconjunto de p = 1,…., k variables. El R_a^2 no ajustado es creciente a medida que se van agregando variables adicionales y, eventualmente, alcanza a valer la unidad cuando el número de parámetros es igual al número de observaciones. El ajustado introduce una penalización por cada parámetro estimado dividiendo SSE y SSTO por sus grados de libertad. Así, puede suceder que disminuya cuando se le introducen variables adicionales al modelo. Así se puede seleccionar el subconjunto particular de regresión para el cual es máximo o para el cual S^2 es mínimo.

Un estadístico alterno para la selección del modelo es C_p sugerido por Mallows (1973), el cual mide la imparcialidad en el modelo de regresión y es de la forma:

$$C_p = \left[\frac{SSE_p}{s^2}\right] - (n - 2_p)[14]$$

En esta expresión, SSE_p es la suma residual de cuadrados de un modelo que contenga p parámetros (incluyendo el intercepto β_0; así, p = 1,…., k+l), y S^2 es el error medio cuadrático del modelo con k variables independientes. Aquí se supone que S^2 es un estimador imparcial de σ^2. Entonces, si una regresión con sólo p parámetros es adecuada, E [SSE_p] = (n - p) σ^2, y aproximadamente, E[C_p] = p. Un gráfico de C_p contra p indica lo adecuado de los modelos a medida que sus puntos se aproximan a la línea C_p = p. Se trata entonces de buscar modelos con un C_p bajo que sea muy aproximado a p.

IX. MULTICOLINEALIDAD

En muchos casos donde se aplica la regresión se encuentra que las variables predictoras X_1,…., X_p están altamente relacionadas mostrando las columnas de la matriz X casi linealmente dependientes. Tal situación usualmente descrita como multicolinealidad en las variables predictoras, es especialmente común si los modelos se ajustan a datos observados. En aquellas situaciones donde el analista puede diseñar el experimento (p. ej. escoger los niveles de las variables independientes, la multicolinealidad puede ser evitada, puesto que siempre se pueden elegir las variables independientes ortogonales ($\sum x_{t-i}x_{t-j} = 0\ para\ i \neq j$).

Sin embargo, cuando se trabaja con observaciones no se puede elegir "la matriz del diseño". Aún más, con demasiada frecuencia, varias variables predictoras miden el mismo concepto. Por ejemplo, un conjunto de indicadores que miden "el estado de la economía", el ingreso familiar y los activos que miden "riqueza", las ventas y el número de empleados que miden 'tamaño del negocio", y la calificación promedio y el puntaje en pruebas estándar que miden "potencial académico" de un aspirante. Todas estas variables tienden a ser altamente relacionadas (correlacionadas). Consecuentemente, la matriz X'X está "mal condicionada" en el sentido de que su determinante es muy pequeño. Esto a su vez crea dificultades de cálculo cuando se invierte esta

matriz para obtener los estimadores mínimos cuadráticos.

Para ilustrar más claramente las consecuencias de la multicolinealidad considérese un ejemplo bien simple, aunque hipotético. Supóngase que se estudia la relación entre el peso de unos automóviles y su consumo de combustible en kilómetros (Y). Sea que se ha medido el peso tanto en libras (X_1) como en kilogramos (X_2). Obviamente no hay información en X_2 puesto que libras y kilogramos guardan una relación exacta de proporcionalidad. Ahora, supóngase que se elige ajustar un modelo de la forma $Y = \beta_0+\beta_1x_1+\beta_2x_2+E$ a una muestra de tamaño n. Aquí encontramos que no se pueden estimar los parámetros puesto que la matriz 3 x 3, X'X, es de rango 2 y no existe su inversa.

Si se intenta entrar los datos a un computador, se detiene el programa y debe aparecer un mensaje de error indicando una sobreinformación en la inversión de la matriz. Por lo tanto, no se pueden estimar los parámetros. Pero si se considera la relación proporcional entre X_1 y X_2 (es decir $X_2=cX_1$) se puede expresar el modelo como $Y = \beta_0+(\beta_1 +c\ \beta_2\ )X_1 +E= \beta_0+\beta' X_1 + E$; de esta manera se puede estimar el parámetro β'. Sin embargo, hay un número infinito de valores para β_1 y β_2 que satisfacen $\beta' = \beta_1 +c\ \beta_2$; por lo tanto, los parámetros individuales β_1 y β_2 no son estimables.

En la mayoría de los casos las relaciones lineales entre las variables independientes no son tan exactas como se supuso en el párrafo anterior, sino que son aproximadas. Por lo tanto, aunque exista la inversa $(X'X)^{-1}$, su determinante puede ser tan pequeño que su obtención no dejará de tener dificultades.

Las consecuencias de una matriz X'X "mal condicionada" son: la diagonal de su inversa muy grande y por lo tanto varianzas grandes para los estimadores mínimos cuadráticos, y altas correlaciones entre los estimadores de los parámetros. Aún más, debido a la gran incertidumbre, se puede presentar mucha inestabilidad en estos estimadores, pues pueden tener el signo equivocado y sugerir, más de lo normal, consideraciones prácticas o físicas.

X. VARIABLES INDICADORAS

Hasta ahora se ha supuesto que las variables independientes son cuantitativas y medidas en una escala bien definida. Sin embargo, con mucha frecuencia las variables son cualitativas y tienen dos o más niveles. Por ejemplo, considérese la predicción del resultado de una cierta reacción química en términos de la concentración de los insumos, la presión, el tiempo de reacción (todas variables cuantitativas), y del tipo de catalizador, (cualitativo: tipo 1,...., tipo k).
En todos estos ejemplos se observa una variable cualitativa a diferentes niveles. Para modelar su efecto se introducen variables adicionales. El efecto de una variable cualitativa que se observa a k niveles diferentes (como si fuera de k fábricas diferentes) en la respuesta de Y, puede representarse por k-1 variables indicadoras. Estas variables indicadoras o ficticias se definen como IND_{it} = 1 si las observaciones vienen del nivel i (para $1 \le i \le k-1$) y 0 en otro caso. Combinando los efectos de estos indicadores con los efectos de p variables cuantitativas X_1, …., X_p, se puede expresar el modelo como:

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i x_{ti} + \sum_{i=1}^{k-1} \sigma_i IND_{ti} + E_t \qquad [15]$$

Si las observaciones varían desde el nivel k, el efecto de las variables cuantitativas independientes está dado por:

$$E(y_t) = \beta_0 + \sum_{i=1}^{p} \beta_i x_{ti} \qquad [16]$$

Si las observaciones varían desde et nivel i ($1 \le i \le k-1$), su efecto es:

$$E(y_t) = \beta_0 + \delta_i \sum_{i=1}^{p} \beta_i x_{ti} \qquad [17]$$

donde el parámetro es el efecto del nivel i relativo al nivel k. Este efecto es aditivo y no

depende de las otras variables independientes $X_1, \ldots, X_p$.

La representación gráfica de la variable indicadora no es única. Alternativamente, se puede definir $IND_{ti} = 1$ si la observación desde el nivel i + 1 (para $1 \leq i \leq k - 1$), y 0 en otro caso. En este caso el parámetro representa los efectos de los niveles comparados con el nivel 1. Alternativamente, se pudieron definir k indicadores y omitir el intercepto β_0. Aquí, los k parámetros δ_1 representan los interceptos de k planos paralelos p-dimensionales.

Los parámetros son fácilmente estimables para cualquier representación que se adopte. Las columnas de la matriz **X** del modelo se componen de los valores de las variables independientes x_{ti} y las demás por columnas unos y ceros.

En el primer caso visto aquí los niveles de la variable cualitativa no interactúan con las otras variables en el modelo. Para considerar las interacciones se pueden introducir términos de productos cruzados $x_{ti}\ IND_{tj}$ ($1 \leq i \leq p$ y $1 \leq j \leq k - 1$). Por ejemplo, el modelo adecuado de producto cruzado para p = 1 y k:

$$y_t = \beta_0 + \beta_1 x_t + \delta_1 IND_{t1} + \delta_2 IND_{t2} + \delta_{11} x_t IND_{t1} + \delta_{22} x_{t1} IND_{t2} + E_t \qquad [18]$$

Las variables indicadoras también pueden usarse para estimar segmentos en modelos de regresión lineal. Considérese un modelo de regresión lineal con una variable predictora, p = 1. Suponga, por ejemplo, que las ventas y_t crecen linealmente con la inversión en publicidad: $y_t = \beta_0 + \beta_1 x_t + E_t$.

Sin embargo, suponga que, si el costo de la publicidad excede una cierta cantidad, X', el efecto de cada peso adicional gastado será menos de β_1. Entonces esta regresión lineal marginada puede ser modelada como:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 (x_t - x^*) IND_t + E_t \qquad [19]$$

donde $IND_t = 1$ si $x_t \neq x'$ y el parámetro β_2 mide la pendiente.

XI. PRINCIPIOS ESTADÍSTICOS GENERALES PARA LA CONSTRUCCIÓN DE MODELOS

Se ha supuesto que se conoce la función del modelo de regresión y la aplicabilidad de los estimadores mínimos cuadráticos y los pronósticos de error medio mínimo cuadrático. Estas propiedades dependen de las hipótesis del modelo, tales como independencia, igual varianza y normalidad de los términos del error.

La inferencia estadística es una herramienta invaluable para el análisis de los datos y la construcción de modelos, pero por ninguna razón es la única. Previamente a la estimación de parámetros se debe especificar la relación funcional entre las variables dependientes e independientes. Luego de la estimación de los parámetros se debe evaluar el ajuste del modelo. Se debe determinar lo adecuado de la relación funcional supuesta, y si se satisfacen las hipótesis de normalidad, igualdad de varianzas e independencia de los errores. Si el modelo ajustado es inadecuado, se debe plantear uno nuevo y se debe repetir el ciclo de estimación y verificación. Solamente cuando el modelo pasa estas revisiones, puede ser usado para interpretación y pronóstico.

Un principio muy importante en la construcción de modelos es el de la parsimonia, también llamado la barbera de Ockham, el cual se puede interpretar como si "en la elección de algunos modelos competitivos, todo lo demás igual, se prefiere el más simple". Preferir los modelos simples a los de más parámetros se debe a que son más fáciles de explicar e interpretar y la explicación de cada parámetro innecesario aumenta la varianza del error de predicción en 1/n. Esto se comprueba al notar que la matriz de covarianzas del vector de valores ajustados $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'$ y está dado por $V(\hat{y}) = \sigma^2(X'X)^{-1}X'$. La varianza promedio de un valor ajustado está dado por:

$$\frac{1}{n}\sum_{t=1}^{n} V(\hat{y_t}) = \frac{\sigma^2}{n} Tr[X(X'X) - 1X'] = p\frac{\sigma^2}{n} \qquad [20]$$

donde Tr(A) es la traza de una matriz cuadrada A, o sea la suma de los elementos de su diagonal, y $\sigma^2 (1 + p/n)$ es la varianza promedio del error de los pronósticos.

XII. CONCLUSIONES

A través del software existente en la actualidad se puede comprobar fácilmente lo adecuado del ajuste del modelo elegido. Un modelo puede ser inadecuado por varias razones:

a. La forma funcional puede ser incorrecta, y
b. La especificación del error puede que no sea adecuada. En particular, puede que los errores no se distribuyan normalmente, o haya desigualdad en las varianzas del error, o que los errores estén correlacionados.

Se define un residual como $e_t = y_t - y_t^{\wedge}$, para t = 1, 2, 3, ... donde y_t es una observación y $y_t^{\wedge}$ es el valor ajustado correspondiente obtenido del análisis de regresión. El análisis de los residuales es muy útil para verificar la hipótesis de la distribución normal $(0, \sigma^2)$ de los errores y para determinar si son útiles los términos adicionales incluidos en el modelo. Los residuales e_t no son variables aleatorias independientes porque involucran los valores ajustados $y_t^{\wedge}$ los cuales se basan en los estimadores muestrales b_0 y b_1. Así, los residuales se asocian con sólo n-2 grados de libertad. Cuando el tamaño muestral es grande en comparación con el número de parámetros en el modelo de regresión, la dependencia entre los e_t pierde importancia y puede ser ignorada. Un gráfico de los residuales contra la variable independiente es supremamente útil para ver si una función de regresión lineal es la adecuada y para comprobar si la varianza de los términos del error es constante.

Se puede considerar el uso de los residuales para examinar seis tipos importantes de desviaciones del modelo aplicado al cual se le asignan errores normales:

1. La función de regresión no es lineal.
2. Los términos del error no tienen varianza constante.
3. Los términos del error no son independientes.
4. El modelo se ajusta a unas observaciones atípicas.
5. Los términos del error no se distribuyen normalmente.
6. Una o varias variables independientes importantes han sido omitidas en el modelo.

En la verificación del diagnóstico, se observan los residuales $e_t = y_t - y_t^{\wedge}$ puesto que expresan la variación que el modelo de regresión no pudo explicar. Se puede considerar que los residuales e_t son los estimadores de E_t; de esta forma, si el modelo ajustado es correcto, los residuales deben confirmar las hipótesis formuladas acerca de los términos del error. Los diagramas de puntos (o histograma) de los residuales e_t contra los valores ajustados $y_t^{\wedge}$, o contra cada variable independiente o contra el tiempo (si los datos fueron recogidos cronológicamente), son de especial utilidad en esta etapa del análisis.

Si se satisfacen las hipótesis del modelo, el gráfico debe semejar una distribución normal. Al graficar los residuales contra los valores ajustados, respecto de cada variable independiente, o contra el tiempo, deben variar en una banda horizontal alrededor de cero. Cualquier desvío se tomará como una indicación de lo inadecuado del modelo.

BIBLIOGRAFIA

Box, G. E. P., and G. C. Tiao, Intervention Analysis with Applications to Economic and Environmental Problems, Journal of the American Statistical Association March 1975, V (70) 34. p.p. 7079.

Chang, lh, George C. Tiao, Estimation of Time Series Parameters in the Presence of Outliers. Technical Report # 8 S.F.

Harvey, A. C., S. Peters, Estimation Procedures for Structural Time Series Models. Journal of Forcasting, Vol 9, 89108 (1990).

Liu, Lon-Mu, and Dominique M. Hanssens, Identification of Transfer Function Model Via Least Squares, Technical Report # 68, Department of Biometrics, UCLA.

RECEPCION: 10/02//2019
APROBACION: 05/03/2019