CICLOS ECONÓMICOS DE PUERTO RICO: USO DE MODELOS Y TÉCNICAS ESTADÍSTICAS MULTIVARIADAS

Por:

John Villavicencio Mattos

Tesis sometida en cumplimiento parcial de los requisitos para el grado de

MAESTRO EN CIENCIAS

en

MATEMÁTICAS (ESTADÏSTICA)

UNIVERSIDAD DE PUERTO RICO

RECINTO UNIVERSITARIO DE MAYAGÜEZ

2010

Aprobado por:

---------- ----------

Julio Quintana Díaz, Ph.D. Fecha

Presidente, Comité Graduado

---------- ----------

Edgardo Lorenzo, Ph.D. Fecha

Miembro, Comité Graduado

---------- ----------

José Alameda, Ph.D. Fecha

Miembro, Comité Graduado

---------- ----------

Silvestre Colón , Prof. Fecha

Director Interino,

Departamento de Ciencias Matemáticas

---------- ----------

Shawn Hunt, Ph.D. Fecha

Representante de Estudios Graduados

ABSTRACT

Economic cycles provide the best monthly economic indicators of a country, both on the long and short term scale. For this reason it is necessary that the cycles model the economy to the most optimal degree possible. This research project proposes a new model for the economic cycles in Puerto Rico. This model, based on the findings of our study, works better than the previous system because it eliminates the problems of multicollinearity and autocorrelation.

This study does a comparative analysis of different techniques such as Discriminant Analysis and Logistic Regression to examine which of these methods better predicts economic cycles. According to the Discriminant Analysis method the variables "Motor Vehicle" and "Salary of the Manufacturer" are those that contribute significantly to discrimination of other groups, while with Logistic Regression we found that the variables "Motor Vehicle", "Sale of Cement" and "Salary of the Manufacturer" were those that contributed to the discrimination.

In conclusion, this study would seem to indicate that the Discriminant Analysis method is preferable to the other because of the fact that this technique results in a lower rate of poor classification.

RESUMEN

Los ciclos económicos son los mejores indicadores económicos mensuales, tanto a corto y largo plazo para un País, por ello se requiere que estos expliquen la economía de manera óptima, es por ello que en este trabajo de investigación se plantea un nuevo modelo para estudiar los ciclos económicos de Puerto Rico, modelo que no presenta problemas de multicolinealidad y autocorrelación.

En el presente trabajo se hizo análisis comparativo de las técnicas multivariadas como Análisis Discrimante y Regresión Logística para ver cuál de estos métodos predice mejor los Ciclos Económicos de Puerto Rico. Según el Análisis Discriminante las variables Vehículo de Motor y Nómina de Manufactura, son las que contribuyen de manera significativa en la discriminación de grupos, mientras que en la Regresión Logística las variables Vehículo de Motor, Venta de Cemento y Nómina de Manufactura, son las que contribuyen de manera significativa en la discriminación de grupos.

El Análisis Discriminante resultó ser mejor ya que esta técnica arroja una menor tasa de mala clasificación.

DEDICATORIA

A mis padres Zenón y Agripina por todo su amor y cariño
A mi esposa por su amor ternura y comprensión
A mis hermanos por su apoyo incondicional

AGRADECIMIENTOS

A Dios, por su apoyo en momentos de soledad y adversidad.

Al Dr. Julio C. Quintana, Presidente del Comité Graduado, por su apoyo en la presente tesis.

Al Dr. José I. Alameda Lozada, por su apoyo incondicional en el presente trabajo de investigación.

A la Junta de Planificación Económica de Puerto Rico, por su apoyo con la facilitación de los datos.

Al Departamento de Ciencias Matemáticas del Recinto Universitario de Mayagüez, por el apoyo brindado para cursar mis estudios de Maestría.

A mis amigos del Departamento de Matemáticas.

LISTA DE TABLAS

Página

LISTA DE FIGURAS

Página

TABLA DE CONTENIDO

Página

1.-INTRODUCCIÓN

1.1 Justificación:

Cuando se analizan los ciclos económicos de Puerto Rico es necesario conocer y estudiar las fluctuaciones de la economía a corto y mediano plazo, quien nos ayuda a estudiar esto es el Índice de Actividad Económica, la Junta de Planificación Económica de Puerto Rico calcula este índice con las siguientes variables: Empleo Total, Índice en Manufactura, Índice de Construcción, Índice de Turismo, Índice de Comercio Exterior, Producción de Energía Eléctrica, Vehículo de Motor y Ventas al Detal. Es necesario conocer las interrelaciones que existen entre estas variables y el efecto de éstas sobre la economía y los ciclos económicos. Después de que se analizan estas variables se necesita la formulación de modelos matemáticos predictivos como el Análisis Discriminante y Regresión Logística, que nos permitan hacer pronósticos sobre el estado de la economía de Puerto Rico a corto y mediano plazo.
La aplicación de estos modelos es fundamental para la toma de decisiones y la formulación de política económica para la Junta de Planificación de Puerto Rico.

Al analizar los ciclos económicos y el Índice de Actividad Económica (IAEJP) se pueden considerar dos posiciones: por un lado, el análisis y medición de los ciclos económicos y sus predicciones y por otra parte estudiar las variables que son más influyentes en la economía de Puerto Rico.

Son estas las razones que nos llevan a analizar, medir y estudiar los ciclos económicos y el Índice de Actividad Económica de Puerto Rico y proponer un modelo para realizar pronósticos a corto y mediano plazo sobre la economía de la Isla.

1.2 Objetivos

- Estudiar las relaciones que existen entre los indicadores económicos de Puerto Rico en periodos de recesión y expansión.
- Modelar los ciclos económicos de Puerto Rico y predecir los futuros cambios que podría experimentar la economía del país a corto y mediano plazo.
- Comparar los métodos del Análisis Discriminante Lineal y de Regresión Logística en el estudio de los ciclos económicos de Puerto Rico.
- Proponer un Índice de Actividad Económica nuevo**,** que difiere del desarrollado por la Junta de Planificación en términos del proceso de selección de variables que influyen en ese índice.
- Determinar cuáles de las variables son las que más influyen en el Índice de Actividad Económica**.**
- Encontrar un modelo matemático para los ciclos económicos de Puerto Rico**.**

2. REVISIÓN DE LITERATURA

2.1. Análisis Discriminante

El Análisis Discriminante es una técnica estadística multivariante de clasificación de elementos en la que se presupone la existencia de dos o más grupos bien definidos a priori.

Los objetivos básicos son:

- Identificar qué variables son las que mejor discriminan entre los grupos.
- Construir una regla de decisión que asigne a un individuo nuevo, que no se sabe clasificar previamente, a uno de los grupos definidos a priori, con cierto grado de riesgo.

Esta técnica de clasificación consiste en obtener unas funciones lineales de las variables independientes, denominadas funciones discriminantes, que permitan clasificar a los individuos en uno de los grupos establecidos por los valores de la variable dependiente.

La base del Análisis Discriminante consiste de una base de datos en los valores de n observaciones en p variables cuantitativas $X^1, X^2, X^3, \dots, X^p$. Los n casos están agrupados en g grupos establecidos por los valores de una variable dependiente U: $G_1, G_2, \dots, G_g$. El grupo G_i consiste en n_i casos ($i = 1, \dots, g$). Por tanto, $n = \sum_{i=1}^{g} n_i$. La tabla de datos establecida por las observaciones tendrá n filas y p columnas. Cada fila puede ser considerada como un punto en un espacio de p dimensiones donde las coordenadas de cada punto se obtendrán a partir de los valores en las p variables para el elemento correspondiente: $x_{ij} = \left(x_{ij}^1, x_{ij}^2, \dots, x_{ij}^p\right) \in R^p,\ i = 1, \dots, g,\ j = 1, \dots n_i$. A partir de la representación de las n filas se trata de extraer un nuevo espacio de pequeña dimensión tal que, al proyectar la nube de puntos sobre dicho espacio, por un lado, los puntos correspondientes a elementos del mismo grupo estén próximos y, por otro, los correspondientes a elementos de los otros grupos distintos estén alejados. Los ejes de este nuevo espacio se llamarán $funciones\ discriminantes$. La expresión de una función discriminante está dada por:

$$Y^k = a_{k1}X^1 + a_{k2}X^2 + \cdots + a_{kp}X^p \quad k = 1,2, \dots, s$$

donde se supone que se extraen s funciones discriminantes. Las puntuaciones discriminantes para el elemento j del grupo G_i estarían dadas por:

$$y_{ij}^{k} = a_{k1}x_{ij}^{1} + a_{k2}x_{ij}^{2} + \cdots + a_{kp}x_{ij}^{p} \quad k = 1,2,\ldots,s$$

A partir de las puntuaciones discriminantes, un elemento para el que se conoce a cuál de los grupos pertenece, será clasificado en uno de ellos. El porcentaje de casos correctamente clasificados será un índice de la efectividad de las funciones discriminantes. Si dichas funciones son efectivas sobre la muestra observada, es de esperar que también lo sean cuando se trate de clasificar a un elemento para el que se desconoce aún a cuál de los grupos pertenece, [Klijn, (2001)].

2.1.1. Clasificación de los datos

Una vez encontradas las funciones discriminantes, es decir, las combinaciones lineales de las variables independientes, a cada dato se le puede asignar una puntuación en la función discriminante.

La discriminación lineal se basa en el siguiente hecho: Un objeto X es asignado al grupo g_1 si

$$D(X, g_1) < D(X, g_2) \qquad (1)$$

donde $D(X, g_i) = (X - U_i)^T \Sigma^{-1}(X - U_i), para\ i = 1,2,$ representa el cuadrado de la distancia Mahalanobis entre x y el centro del grupo g_i.

Realizando operaciones algebraicas de matrices, la desigualdad (1) se puede escribir como

$$(U_1 - U_2)^T \Sigma^{-1} \left(X - \frac{1}{2}(U_1 - U_2) \right) > 0$$

Para una muestra, $\overline{X_i}$ estima a U_i, y S estima a Σ, la matriz de covarianza muestral combinada, la cual se calcula por:

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

2.1.2. La Regla de Bayes

Se pueden usar las puntuaciones discriminantes para obtener una regla para clasificar los casos en los grupos.

Así, la probabilidad de que un objeto, con una puntuación discriminante D pertenezca al grupo $i - ésimo$ se puede estimar mediante la regla de Bayes:

$$P(G_i/D) = \frac{P(D/G_i)P(G_i)}{\sum_{i=1}^{g} P(D/G_i)P(G_i)}$$

donde:

$P(G_i)$ es la probabilidad a priori y es un estimado de la confianza de que un objeto pertenezca a un grupo si no se tiene información previa.

$P(D/G_i)$ es la probabilidad de obtener la puntuación D estando en el grupo $i - ésimo$

$P(G_i/D)$ es la probabilidad de que un objeto pertenezca al grupo G_i, dado que presenta la puntuación D, [Cuevas y Berrendero, (2003)].

2.1.3 Tasa de mala clasificación

La tasa de mala clasificación es la probabilidad de que el clasificador clasifique mal una observación de la población a la cual pertenece la muestra usada para construir el clasificador. Existen varios métodos de estimar la tasa de error de clasificación. A continuación se describen dos métodos.

- **Error aparente:** Es conocido también como el criterio de estimación de la tasa de error por resustitución [Smith. (1947)]; aquí se utiliza toda la muestra, luego se clasifican estas mismas observaciones y por comparación con su verdadera clase se obtiene una proporción de observaciones mal clasificadas.

  Este criterio no es muy recomendado, pues subestima el error de mala clasificación y se puede llegar a falsas conclusiones si el tamaño de la muestra no es muy grande al comparar con el número de variables utilizadas en el modelo, [Acuña, (2005)].

- **Tasa de error por validación cruzada:** Este método fue propuesto por Stone (1974) que consiste en dividir la muestra en m subconjuntos excluyentes de igual tamaño $n_1, .., n_i, \ldots .., n_m$. Usualmente se considera $m = 10$ para estimar el modelo de clasificación usando todas menos una de las submuestras; luego se clasifican las observaciones que se dejaron de lado; el promedio de las clasificaciones erradas dará el estimado de la tasa de error por validación cruzada. Este método define un estimador con poco sesgo, [Acuña, (2005)].

2.2. Regresión Logística

Inicialmente la regresión logística fue sugerida por Cox (1970), la condición de la existencia de una única solución para la ecuación de verosimilitud fue dada por Albert y Anderson (1984). La regresión logística es una técnica estadística multivariante que

permite estimar la relación existente entre una variable dependiente de tipo cualitativo, en caso particular de tipo binario, y un conjunto de variables explicativas, éstas pueden ser cuantitativas o cualitativas. Se considera una muestra de tamaño $n = n_1 + n_2$, donde n_1 observaciones son de la clase uno, mientras que n_2 son de la clase dos. Así, para cualquier observación x_j la variable de respuesta Y es igual a 1 si x_j es de la clase uno, mientras que Y es igual a 0 si x_j pertenece a la clase dos.

La Regresión Logística es equivalente al Análisis Discriminante. Este último es óptimo para clasificar si la distribución conjunta de las variables explicativas es normal multivariante; con varianzas iguales e igual matriz de covarianzas. Sin embargo, la discriminación lineal puede funcionar mal en otros contextos, cuando las covarianzas son distintas o las distribuciones difieren mucho de la normal. En estos casos el modelo de regresión logística puede conducir a mejores resultados [Vicente, (2006)].

Cuando el modelo de regresión logística tiene una sola variable explicativa, que además es binaria, entonces existe una relación entre la regresión logística y el análisis de una tabla de contingencia 2x2.

Los objetivos básicos de la regresión logística son:

- Construir un modelo que permita asignar el valor de la variable dependiente (probabilidad del suceso) para unos valores determinados de un conjunto de variables explicativas.
- Determinar el modelo más parsimonioso y mejor ajustado que siendo razonable describa la relación entre la variable respuesta y un conjunto de variables explicativas.

2.2.1. Regresión logística simple

En el modelo de regresión logística $E(y_i/x_i) = \beta_o + \beta_1 x_i$ con $y_i = 0{,}1$; que representa la esperanza condicional de y_i dado x_i, en forma específica la distribución acumulada de la distribución logística está dada por

$$\pi_i = \frac{e^{\beta_o+\beta_1 x_i}}{1+e^{\beta_o+\beta_1 x_i}} \qquad (2)$$

Una transformación de π_i es de suma importancia en el modelo de regresión logística. Esta transformación se define en términos π_i.

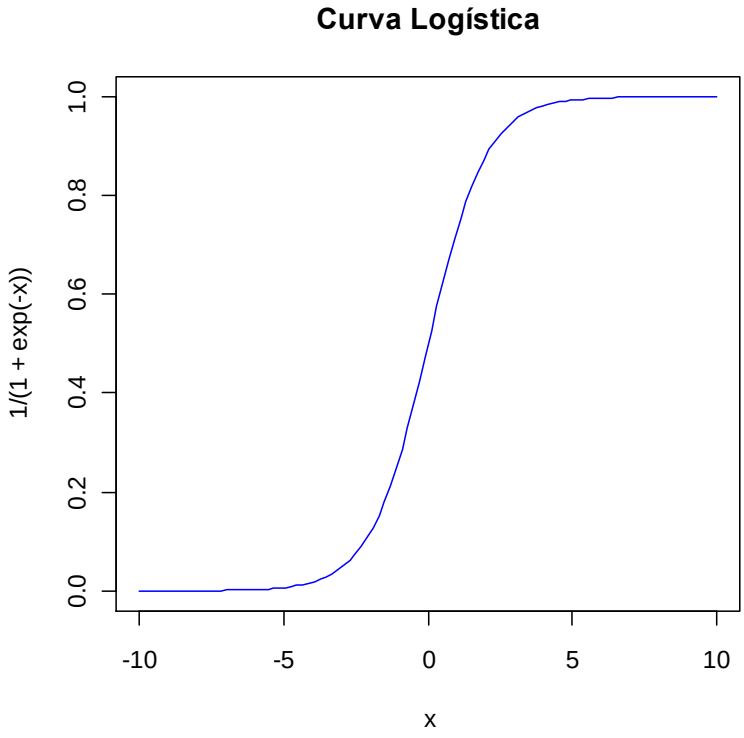
$$g(x_i) = ln\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_o + \beta_1 x_i \qquad (3)$$

La importancia de esta transformación es que $g(x_i)$ tiene muchas propiedades de un modelo de regresión lineal. La función $g(x_i)$ es lineal en sus parámetros, es continua y puede tomar valores en $(-\infty, +\infty)$.

Cuando se trabaja en regresión lineal se tiene que $y_i = E(y_i/x_i) + \varepsilon_i$, donde ε_i identifica el error aleatorio o perturbación de la i-ésima observación ocasionada por variables no medidas. En este modelo se tienen que cumplir algunos supuestos como los de normalidad, independencia y varianza constante de los errores. Este no es el caso cuando la variable dependiente es dicotómica. Cuando esto ocurre el modelo queda como $y_i = \pi_i + \varepsilon_i$; donde el valor de ε_i toma una de dos posibilidades. Si $y_i = 1$ entonces $\varepsilon_i = 1 - \pi_i$ con probabilidad π_i, y si $y_i = 0$ entonces $\varepsilon_i = -\pi_i$ con probabilidad $1 - \pi_i$. Así ε_i tiene una distribución con media cero y varianza $\pi_i[1 - \pi_i]$, Por tanto, la distribución condicional de la variable respuesta tiene distribución binomial con media π_i .

Se espera una relación curvilínea de π_i y x_i. Para cualquier valor "grande" de x_i, π_i tomará valores "grandes" cercanos a 1 y para valores "pequeños" de x_i, π_i tomará valores cercanos a cero. Una gráfica del modelo logístico se muestra a continuación.

FIGURA 1



Esta gráfica en forma sigmoidal tiene las propiedades requeridas para π_i y además tiene las propiedades de una función de distribución de probabilidad acumulada.

En la ecuación (2) cuando $\beta_0 < 0$ y $\beta_1 > 0$, la gráfica es similar a la gráfica 1.

Cuando $P[y_i = 1] = 0.5$, el valor de x_i es igual a $-\frac{\beta_o}{\beta_1}$.

Una de las características que hace tan interesante la regresión logística es la relación que estos parámetros guardan con un parámetro de cuantificación de riesgo conocido como "Odds Ratio"o razón de probabilidad de $y_i = 1$ contra $y_i = 0$, específicamente ese parámetro está dado por: $OR = \frac{\pi_i}{1-\pi_i}$, llamándosele también "razón de ventaja a favor del éxito".

2.2.1.1. Estimación de los parámetros del modelo logístico simple

Dado que cada observación y_i es una variable aleatoria Bernoulli, donde

$P(y_i = 1) = \pi_i$ y $P(y_i = 0) = 1 - \pi_i$

Entonces la distribución de probabilidad de la variable Y se puede representar por: $f_i(y_i) = {\pi_i}^{y_i}(1 - \pi_i)^{1-y_i}$ donde $y_i = 0{,}1;\ \ i = 1{,}2{,}\ldots, n.$ Además $f_i(1) = \pi_i$ y $f_i(0) = 1 - \pi_i$. Por lo tanto $f_i(y_i)$ simplemente representa la probabilidad de que $y_i = 1$ ó 0.

Como las y_i son observaciones independientes, entonces la función de distribución conjunta es:

$$g(y_1, \ldots, y_n) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} {\pi_i}^{y_i}(1 - \pi_i)^{1-y_i}$$

Será más fácil encontrar las estimaciones de máxima verosimilitud mediante el logaritmo de la función de distribución conjunta.

$$ln\big(g(y_1, \ldots, y_n)\big) = ln\left(\prod_{i=1}^{n} {\pi_i}^{y_i}(1 - \pi_i)^{1-y_i}\right)$$

$$ln\big(g(y_1, \ldots, y_n)\big) = \sum_{i=1}^{n}\left[y_i ln\left(\frac{\pi_i}{1 - \pi_i}\right)\right] + \sum_{i=1}^{n} ln(1 - \pi_i) \qquad (4)$$

De la ecuación (2), π_i se puede expresar de la siguiente forma $\left[1 + e^{-\beta_o - \beta_1 x_i}\right]^{-1}$, por lo que:

$$1 - \pi_i = \left[1 + e^{\beta_o + \beta_1 x_i}\right]^{-1} \qquad (5)$$

En la ecuación (4) se sustituyen las ecuaciones (3) y (5)

$$L(\beta) = L(\beta_o, \beta_1) = \sum_{i=1}^{n} y_i(\beta_o + \beta_1 x_i) - \sum_{i=1}^{n} ln\left[1 + e^{\beta_o + \beta_1 x_i}\right] \qquad (6)$$

Para encontrar el valor de β que maximice $L(\beta)$ encontramos las derivadas parciales de $L(\beta)$ con respecto a β_o y β_1, las ecuaciones de verosimilitud que se obtienen se igualan a cero.

$$\frac{\partial L(\beta_o, \beta_1)}{\partial \beta_o} = \sum_{i=1}^{n} \left(y_i - \frac{e^{\beta_o + \beta_1 x_i}}{1 + e^{\beta_o + \beta_1 x_i}} \right) = \sum_{i=1}^{n} (y_i - \pi_i) = 0 \qquad (7)$$

$$\frac{\partial L(\beta_o, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} \left(y_i x_i - \frac{x_i e^{\beta_o + \beta_1 x_i}}{1 + e^{\beta_o + \beta_1 x_i}} \right) = \sum_{i=1}^{n} x_i (y_i - \pi_i) = 0 \qquad (8)$$

En el caso de la regresión lineal, la función de verosimilitud es más fácil de estimar porque son lineales con respecto a los parámetros. Pero en el caso de regresión logística las expresiones de las ecuaciones (7) y (8) son no lineales en β_o y β_1, y, por lo tanto requieren métodos especiales para su solución. Estos métodos son recurrentes tales como el de método de Newton-Raphson, [Hosmer y Lemeshow, (2000)].

2.2.2. Regresión Logística Múltiple

Es una generalización del modelo de regresión logística simple, donde se considera más de una variable independiente, en donde por lo menos una de ellas es de tipo cuantitativo.

El modelo logístico múltiple está dado por:

$$\pi_i = \frac{e^{\beta_o+\beta_1 x_1+\cdots\ldots+\beta_p x_p}}{1+e^{\beta_o+\beta_1 x_1+\cdots\ldots+\beta_p x_p}} = \frac{e^{\alpha+\beta^T X}}{1+e^{\alpha+\beta^T X}}$$

O también $$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta^T X \qquad (9)$$

Donde X es la matriz de datos, β^Tes la transpuesta del vector de p parámetros y α es el vector que representa el intercepto.

Si las variables X en cada clase se distribuyen normalmente con igual matriz de covarianzas Σ, entonces se satisface la suposición (9) ya que

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = (\mu_1-\mu_2)^T\Sigma^{-1}\left(X - 1/2\,(\mu_1+\mu_2)\right) \qquad (10)$$

En este caso $\alpha = -(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2)/2$ y $\beta = (\mu_1-\mu_2)^T\Sigma^{-1}$. La suposición de (9) se cumple también para otros tipos de distribuciones distintas de la normal multivariada, tales como distribuciones de Bernoulli, y combinaciones de éstas.

$P(Y=1)$ es la probabilidad a posteriori de que Y sea igual a 1 para un valor observado de X. Entonces haciendo uso de la probabilidad condicional, las interpretaciones que se pueden hacer de los coeficientes de la regresión logística son:

- Un coeficiente $\hat{\beta}_k$ en el modelo de regresión logística estimado representa el cambio promedio de la función logit cuando la variable X_k cambia en una unidad adicional suponiendo que las otras variables permanecen constantes.

2.2.2.1 Estimación del modelo logístico múltiple

Para una muestra de $n = n_1 + n_2$ observaciones independientes definido por $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i), con\ i = 1,2, \ldots., n$, con n_1 muestras en la clase 1 y n_2 muestras en la clase 2; como en el caso univariado elegimos el vector de coeficientes $\beta^T = (\beta_0, \beta_1, \ldots, \beta_p)$ y un parámetro binomial

$$\pi_i = \frac{exp\ (\alpha + X_i^T \beta)}{1 + exp\ (\alpha + X_i^T \beta)}$$

A fin de obtener la estimación de máxima verosimilitud para el vector β, la función de densidad de probabilidad está dada por:

$$g(y_1, \ldots, y_n) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} {\pi_i}^{y_i} (1 - \pi_i)^{1-y_i}$$

De donde la función de verosimilitud queda como

$$L(\alpha, \beta) = \prod_{i=1}^{n_1} \frac{exp\ (\alpha + X_i^T \beta)}{1 + exp\ (\alpha + X_i^T \beta)} \cdot \prod_{j=n_1+1}^{n} \frac{1}{1 + exp\ (\alpha + X_j^T \beta)}$$

Tomando el logaritmo natural a $L(\alpha, \beta)$ y derivando parcialmente con respecto al vector β y a α encontramos las $p+1$ ecuaciones de verosimilitud con respecto a las $p+1$ coeficientes. Las ecuaciones de estos resultados son

$$\sum_{i=1}^{n}[y_i - \pi_i] = 0$$

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi_i] = 0 \ , j = 1,2,..,p$$

Los estimados $\hat{\alpha}$ y $\hat{\beta}$ son aquellos que maximizan la función de verosimilitud y se encuentran aplicando métodos recurrentes tales como el de Newton-Raphson en el caso de regresión logística simple.

2.2.3 Medidas de confiabilidad del modelo

Los siguientes métodos son medidas que cuantifican el nivel de ajuste del modelo al conjunto de datos.

2.2.3.1 La Devianza Residual:

Este método fue propuesto por Nelder y Wederbum (1982), es análogo a la suma de cuadrados de los residuales de un modelo de regresión lineal y se define como el negativo de dos veces la función de verosimilitud maximizada.

$$D = -2\sum_{i=1}^{n}\left[y_i ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i) ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right]$$

Donde $\hat{\pi}_i$ es el valor estimado de la ecuación (2). D es equivalente a la prueba de razón de verosimilitud para probar la validez del modelo logístico, [Hosmer y Lemeshow, (2000)].

La prueba de hipótesis es:

$$H_o: El\ modelo\ ajustado\ es\ confiable$$

$$H_a: El\ modelo\ ajustado\ no\ es\ confiable$$

El estadístico de prueba

$$D \sim \chi^2 \ \ con\ n-p-1\ grados\ de\ libertad$$

Donde p es el número de variables regresoras (o explicativas) y n es el número de datos.

Decisión, si $D > \chi^2_{\alpha,n-p-1}$ entonces el modelo logístico no es confiable.

2.2.3.2 El Pseudo-R^2

Se han propuesto versiones similares al R^2 de la regresión lineal. Aquí definimos el propuesto por McFadden [Acuña, (2008)].

$$PseudoR^2 = 1 - \frac{Devianza\ Residual}{Devianza\ Nula}$$

Donde la Devianza Nula es la devianza considerando solamente el intercepto y que se distribuye como una Ji-Cuadrado con $n-1$ grados de libertad. Para hallar la Devianza Nula se hace una regresión logística considerando que hay una sola variable predictora cuyos valores son todos unos. Un Pseudo-R^2 mayor que 0.3 es considerado aceptable, [Acuña, (2008)].

2.2.3.3 El Criterio de Información de Akaike (AIC)

El criterio de AIC (Akaike, 1974) es uno de los más utilizados como medidas de confiabilidad para identificar el mejor modelo.

$$AIC = D + 2(P+1)$$

Donde D es la varinza residual y Pes el número de variables predictoras. Un modelo es mejor que otro si su AIC es más pequeño, [Acuña, (2008)].

2.2.4 Clasificación con regresión logística

Una manera tradicional de discriminar los datos con regresión logística, es considerando que si $\pi_i > 0.5$ entonces la observación pertenece a la clase de interés. Pero algunas veces esta manera de clasificar resulta ser un método heurístico, por no tener un criterio objetivo.

Existen dos métodos alternos para encontrar el π_i óptimo.

- Graficar el porcentaje de observaciones que están en la clase de interés y que han sido correctamente clasificadas (sensitividad) versus distintos niveles de probabilidad y graficar el porcentaje de observaciones de la otra clase que han sido correctamente clasificadas (especificidad) versus los mismos niveles de probabilidad utilizados para la gráfica de sensitividad.
- La curva ROC (Receiver Operating Characteristic). Para encontrar esta curva se grafica la curva de sensitividad versus la curva de (1- especificidad), el π_i óptimo es aquél que está más cerca a la esquina superior izquierda.

2.3. Multicolinealidad

El término de multicolinealidad en Estadística es una situación en la que se presenta una fuerte correlación entre variables explicativas del modelo; esto es, una de las hipótesis del modelo de regresión lineal múltiple establece que no debe existir relación lineal entre las variables explicativas, esto es, no existe *multicolinealidad* en el modelo. Esta hipótesis es necesaria para el cálculo del vector de estimadores mínimos cuadráticos, ya que en caso contrario la matriz $X^T X$ será no singular. En caso de que existe una relación aproximadamente lineal entre las variables regresoras los estimadores que se obtengan serán poco precisos. En otras palabras, la relación entre regresores hace que sea difícil cuantificar con precisión el efecto que cada regresor ejerce sobre la variable dependiente, lo que determina que las varianzas de los estimadores sean elevadas.

Para analizar el problema de multicolinealidad de acuerdo a Besley, et al. (1991) se pueden realizar los siguientes análisis:

2.3.1 Factor de inflación de varianza (VIF)

Este método fue propuesto por Marquardt (1970), en el modelo de regresión lineal múltiple $Y = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p$. La varianza del j-esimo coeficiente de regresión estimado está dada por $Var(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1-R_j^2}\right)\left(\frac{1}{S_{X_j X_j}}\right)$ donde R_j^2 es el coeficiente de determinación de la regresión lineal de X_j versus todas las demás variables explicativas. El número $\frac{1}{1-R_j^2}$ es llamado el j-ésimo FACTOR DE INFLACIÓN DE LA VARIANZA (VIF). Si R_j^2 es cercano a 1 entonces la varianza de $\hat{\beta}_j$ aumentará grandemente. El VIF representa el incremento en la varianza del coeficiente de una variable regresora debido a la presencia de multicolinealidad, [Acuña, (2008)].

Una variable regresora con un VIF mayor a 10 (esto es equivalente a un $R^2 = 0.90$) puede causar multicolinealidad.

2.3.2 Número condición

Este procedimiento de detección de multicolinealidad es uno de los métodos más utilizados entre los actualmente disponibles, según afirman Judge et al (1985).

El número de condición mide la sensibilidad de las estimaciones mínimocuadráticas ante pequeños cambios en los datos. De acuerdo con los estudios realizados por Belsley (1982), el problema de la multicolinealidad es serio cuando el número de condición toma valor entre 20 y 30, si este indicador superase el valor de 30, el problema es muy grave.

Sea U una matriz tal que $Z = XU$ y que $Z^T Z = U^T X^T XU = D$, donde U es una matriz ortogonal y D es una matriz diagonal con elementos positivos $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, los $\lambda_{"s}$ son llamados autovalores de $X^T X$ y las columnas de U son autovectores de $X^T X$. Las columnas de $Z = XU$ son llamados componentes principales, [Wschebor, (2005)].

El número condición K es igual a la raíz cuadrada de la razón entre el autovalor más grande λ_{max} y el autovalor más pequeño λ_{min}, de la matriz $X^T X$ esto es:

$$K = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

Como la matriz $X^T X$ es de dimensión kxk se obtienen k raíces características, pudiéndose calcular para cada una de ellas un índice de condición definido de la siguiente forma:

$$K(\lambda_i) = \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

2.4 Autocorrelación

La autocorrelacion surge cuando los términos de error del modelo no son independientes entre sí, es decir, cuando $E(\mu_i, \mu_j) \neq 0$, para todo $i \neq j$. Entonces los errores están vinculados entre sí. Los modelos de estimadores mínimo cuadráticos ordinarios obtenidos bajo esta circunstancia dejan de ser eficientes, la autocorrelacion generalmente aparece en datos en el tiempo.

El primer problema a resolver es determinar el tipo de autocorrelación, Es decir, debemos caracterizar la forma que adopta la correlación serial de las perturbaciones del modelo. Una primera aproximación, muy utilizada en la práctica, es suponer que la perturbación del modelo sigue un proceso autoregresivo de primer orden.

2.4.1 Modelo autoregresivo de primer orden AR(1)

El modelo autoregresivo de primer orden AR(1) es el caso más común, aquel en el que existe correlación de primer orden entre las perturbaciones del modelo.

El planteamiento a seguir parte de suponer que queremos estudiar la presencia de autocorrelación en el siguiente modelo.

$$Y = X\beta + \mu$$

donde β representa el estimador mínimo cuadrático y $\hat{\mu} = Y - X\hat{\beta}$.

La hipótesis a formular es que las perturbaciones del modelo anterior pueden definirse de acuerdo al siguiente modelo:

$\mu_t = \rho\mu_{t-1} + \varepsilon_t$; para $-1 < \rho < 1$

Este modelo expresa un comportamiento autoregresivo de primer orden de los errores.
A ρ se le conoce como coeficiente de autocovarianza o de autocorrelación, ε_t es un ruido "blanco".
Si $\rho = 0$, entonces $\mu_t \; y \; \mu_{t-1}$ no están correlacionadas, por lo que se concluye que no existe autocorrelación de primer orden.
Si $\rho \neq 0$, entonces $\mu_t \; y \; \mu_{t-1}$ están correlacionadas, y por lo tanto existe problema de autocorrelación en el modelo.
Para contrastar la hipótesis nula $H_o: \rho = 0$ hay diversos estadísticos. De entre ellos, el que se considerará es la prueba de Durbin-Watson, [Montañés, (1995)].

2.4.2 Prueba de Durbin-Watson

Esta prueba fue propuesta por Durbin y Watson (1950). Permite contrastar $H_o: \rho = 0$ vs $H_a: \rho \neq 0$, está define de la siguiente manera

$$DW = \frac{\sum_{i=2}^{T}(\hat{\mu}_t - \hat{\mu}_{t-1})^2}{\sum_{i=2}^{T}\hat{\mu}_t^2} \qquad (11)$$

Donde $\hat{\mu}_t$ es el residuo mínimo cuadrático ordinario del periodo t, con $t = 1,2, \dots, T$

Realizando operaciones en la ecuación (11), el estadístico Durbin-Watson queda

$$DW = 2(1 - \hat{\rho})$$

Donde $\hat{\rho}$ es un estimador consistente del coeficiente de correlación de primer orden.

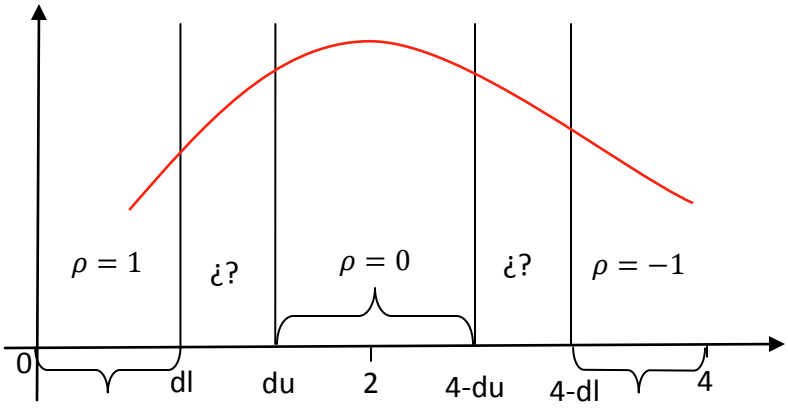
El estadístico de Durbin-Watson no sigue ninguna distribución conocida, la manera de evaluar es bajo las siguientes condiciones.

- Si no existe autocorrelación entre $\mu_t \; y \; \mu_{t-1}$, es de esperar que el valor del estimador del coeficiente de autocorrelación se aproxime a 0, entonces el valor del estadístico DW tome un valor en un entorno a 2.
- Si existe autocorrelación positiva entre $\mu_t \; y \; \mu_{t-1}$, esto es $\hat{\rho}$ toma valores positivos, lo que hará que el estadístico DW se aleje de 2 y se aproxime hacia 0 a medida que el grado de correlación sea mayor.
- Si existe autocorrelación negativa entre $\mu_t \; y \; \mu_{t-1}$, el estadístico de DW se aleja de 2 aproximándose hacia 4 conforme el valor del coeficiente de autocorrelación se acerca hacia -1.

Una manera gráfica de explicar lo anterior es

FIGURA 2

Intervalos del Estadístico Durbin-Watson



Donde dl y du son estadísticos que no depende de la matriz de información X, estos estadísticos dependen del nivel de significancia α, el tamaño de muestra n y el número p de variables en el estudio [Montañés, (1995)].

3. METODOLOGÍA DE OBTENCIÓN DEL ÍNDICE DE ACTIVIDAD ECONÓMICA

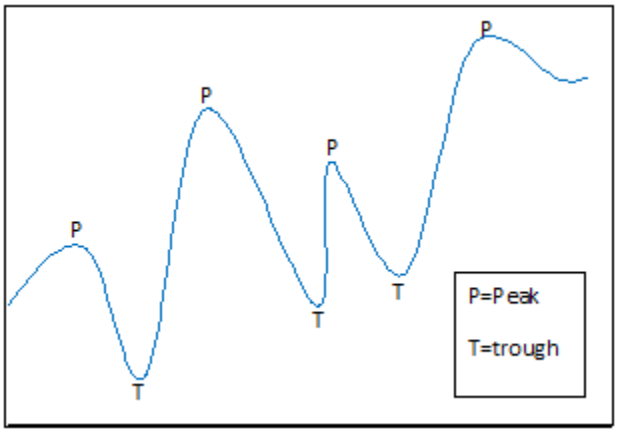
En este capítulo se describe de manera detallada la forma como se calculan los índices de actividad económica de Puerto Rico, para esto se necesita tener conocimiento de las siguientes definiciones, las cuales fueron dadas por Robert Emerson Lucas, publicadas en su libro *Lectures on Economic Growth* (2004)

- **Ciclo Económico.-** Los ciclos son fluctuaciones recurrentes en las actividades económicas. Un ciclo consiste en período de expansión y otro de recesión. Esta sucesión de cambios es recurrente, pero no periódica; la duración del ciclo varía. El único carácter regular de estas fluctuaciones es el modo en que las variables se mueven juntas. Este movimiento conjunto es lo que se llama ciclo económico.
- **Pico (peak).-** El último mes antes de que varios indicadores clave empiezan a caer.
- **Valle (trough).**- El último mes antes de que los mismos indicadores empiezan a crecer.

Para saber si en un determinado ciclo se produjo una recesión o una expansión se aplicará la metodología T-P-T (*trough-peak-trough*), la cual fue propuesta por la Oficina Nacional de Investigación Económica (*National Bureau of Economic Research, NBER*), donde un ciclo se mide desde un punto mínimo (*trough*); luego le sigue un punto máximo (*peak*); y por último termina con punto mínimo (*trough*), que gráficamente se representa así:

FIGURA 3

Ciclos Económicos



Los procedimientos que se siguieron en el presente trabajo de investigación fueron:

a) Como primer paso se recopiló información mensual de las series que componen el Índice de Actividad Económica de la Junta de Planificación (ICAE), desde el año 1971 hasta 2008. Las fechas de referencia fueron tomadas de acuerdo a la información que maneja el Programa de Planificación Económica y Social de la Junta de Planificación.

b) Se verificó si el modelo que maneja la Junta de Planificación Económica tiene problemas de multicolinealidad y autocorrelación.

c) Asociada a los valores del ICAE se definió una variable dicotómica que toma el valor 1 en el caso de que el mes corresponda a un período de expansión y de 0 para un mes de recesión.

d) Se comparó el Análisis Discriminante Lineal y el método de Regresión Logística Múltiple para determinar cuál de estos dos modelos es mejor para hacer pronósticos y además con cuál se obtiene una menor tasa de mala clasificación.

3.1 Pasos para calcular el Índice de Actividad Económica

Para explicar la metodología de cómo se calcula el Índice de Actividad Económica de Puerto Rico, nos referiremos a la revista *Business Cycle Indicators Vol. 1, Number 11, December 1996, The*

Conference Board "Calculating the Composite Indexes". Esta metodología también se utiliza para calcular el Índice de Actividad Económica de los Estados Unidos.

Esta metodología se basa en el cálculo de índices compuestos o coincidentes desarrollada por los investigadores Geoffrey Moore y Julius Shiskin durante los años 50, [Curet,1976] y se describe a continuación:

1. Se calculan los cambios simétricos de las series

$$C_{it} = 200\left(\frac{d_{it} - d_{it-1}}{d_{it} + d_{it-1}}\right)$$

Ó $C_{it} = d_{it} - d_{it-1}$ si d tiene valores negativos o ceros.

donde:

d es el valor de la serie

C_{itt} es el valor del cambio simétrico mensual de la serie.

t es el valor de un mes (su valor va de $t = 1$ hasta $t = T$, el total de valores para la serie).

i es el identificador del componente (su valor va de 1 a 13 en el caso del ICAE).

2. Se calcula el factor de normalización de cada componente del índice.

$$A_i = \frac{\sum|C_{it}|}{N - 1}$$

$\sum|C_{it}|$ es la sumatoria de los valores absolutos de los cambios simétricos.

N es el número de meses u observaciones a ser usadas en el cálculo.

3. Se procede a ajustar los cambios simétricos por el factor de normalización de cada serie. Este procedimiento se hace para evitar los sesgos en las series y que éstas no dominen el movimiento mensual del índice.

$$S_{it} = \frac{C_{it}}{A_i}$$

S_{it} es el factor de normalización de los cambios simétricos de la serie i con base $n - 1$.

4. Una vez calculados los cambios simétricos normalizados para cada una de las series que componen el índice se procede al cómputo de éste bajo las siguientes operaciones:

a. Cómputo de cambios simétricos ya ajustados por factor de normalización de cada serie.

$$R_t = \frac{\sum S_{it}}{N_t}$$

R_t es el factor de normalización del índice.

N_t es el número de componentes o series en ese mes particular.

b. Cómputo del factor de normalización del índice.

$$F = \frac{\sum |R_{it}|}{N - 1}$$

N es el número de observaciones a ser usadas en el cálculo.

c. Cálculo de los cambios simétricos ya ajustados por los factores de normalización de las series con el factor de normalización del índice. Este cálculo es para evitar los sesgos en el índice.

$$r_t = \left(\frac{R_t}{F}\right) 100$$

5. Cálculo del nivel del índice compuesto.

$$I_t = I_{t-1}\left(\frac{200 + r_t}{200 - r_t}\right)$$

I_t es el valor inicial del índice.

Nota: el primer valor es de 100, lo que significa que la fórmula inicial será:

$$I_t = 100\left(\frac{200 + r_t}{200 - r_t}\right)$$

Este valor I_t es multiplicado por 100. De aquí se calcula el valor del año base.

6. Se procede a calcular el índice expresado en un año base. Para ello se calcula el promedio de 12 meses del año base 1960. Una vez obtenido el promedio se divide el valor del índice entre este promedio para obtener el valor del índice sobre la base del año seleccionado.

$$I_{t,67} = \frac{I_t}{X_{67}}$$

Donde:

$$X_{67} = \frac{Suma\ del\ indice\ de\ los\ 12\ meses\ seleccionados\ del\ año\ base}{12}$$

7. La última operación del índice consiste en expresarlo mediante un promedio móvil de 6 meses.

$$IS_{t,67} = \frac{XI_{t,67}}{M_t}$$

Donde:

$$M_{t-1} = \frac{M_{t-1} + M_{t-2} + M_{t-3} + M_{t-4} + M_{t-5} + M_{t-6}}{6}$$

4.- ANÁLISIS EXPLORATORIO

En este capítulo se hará un análisis exhaustivo de la información que maneja la Junta de Planificación Económica de Puerto Rico para calcular el IAE. Este índice está en función de ocho variables que son: Empleo Total, Índice de Manufactura, Índice de Comercio Exterior, Índice de Turismo, Ventas al Detalle, Producción de Energía Eléctrica, Índice de Construcción y Registro de Vehículos de Motor. Esto es:

$$IAE = f(emp, Imf, Ice, Itu, Vad, Pee, Icost, Vmotor)$$

$emp = Empleo$

$Imf = \text{Índice en manufactura}$

$Ice = \text{Índice de comercio exterior}$

$Itu = \text{Índice de turismo}$

$Vad = Ventas\ al\ detalle$

$Pee = Producción\ de\ energía\ electrica$

$Icost = \text{Índice de construcción}$

$Vmotor = Vehículos\ de\ motor$

A su vez los índices de manufactura, comercio exterior, turismo y construcción están en función de otras variables.

$Imf = f(empmf, nommf, hommf)$

$$empmf = Empleo\ en\ manufactura$$

$$nommf = \text{Nómina en manufactura}$$

$$hommf = Horas\ en\ manufactura$$

$Ice = f(expor, import)$

$$expor = Exportaciones$$

$$import = Importaciones$$

$Itu = f(emptur, reghot)$

$$emptur = Empleo\ en\ turismo$$

$$reghot = Registro\ hotelero$$

$Icot = f(empconst, vetcemen, valperconst)$

$$empconst = Empleo\ en\ construcción$$

$$vetcemen = Venta\ de\ cemento$$

$$valperconst = Valor\ de\ permisos\ de\ construcción$$

Para verificar si hay presencia de multicolinealidad se utilizó la matriz de correlaciones.

TABLA 4.1

Matriz de Correlaciones

	emp	*Imf*	*Ice*	*Itu*	*Vmotor*	*Vad*	*Pee*	*Iconst*
emp	1							
Imf	**0.984**	1						
Ice	**0.880**	**0.871**	1					
Itu	**0.972**	**0.989**	0.865	1				
Vmotor	0.660	0.693	0.521	0.702	1			
Vad	**0.885**	**0.917**	0.746	**0.932**	0.727	1		
Pee	**0.930**	**0.951**	0.826	**0.961**	0.723	**0.924**	1	
Iconst	**0.959**	**0.951**	**0.885**	**0.966**	0.646	**0.872**	**0.938**	1

De la tabla 4.1 podemos afirmar que casi todas las variables están altamente correlacionadas. Para corroborar esto se calculó el número condición de la matriz y el factor de inflación de la varianza.

- **Número condición de la matriz**

  NC=36.40091

El número condición de la matriz es mayor de 30, por lo que según se estableció en la sección 2.3.2, las variables están altamente correlacionadas y el problema de multicolinealidad se considera muy grave.

- **Factor de inflación de varianza para cada una de las variables**

También se aplicó el procedimiento descrito en la Sección 2.3.1, el factor de inflación de la varianza (VIF), con los resultados que se muestran en la tabla 4.2:

TABLA 4.2

Factores de Inflación de la Varianza (VIF)
Variables del IAE

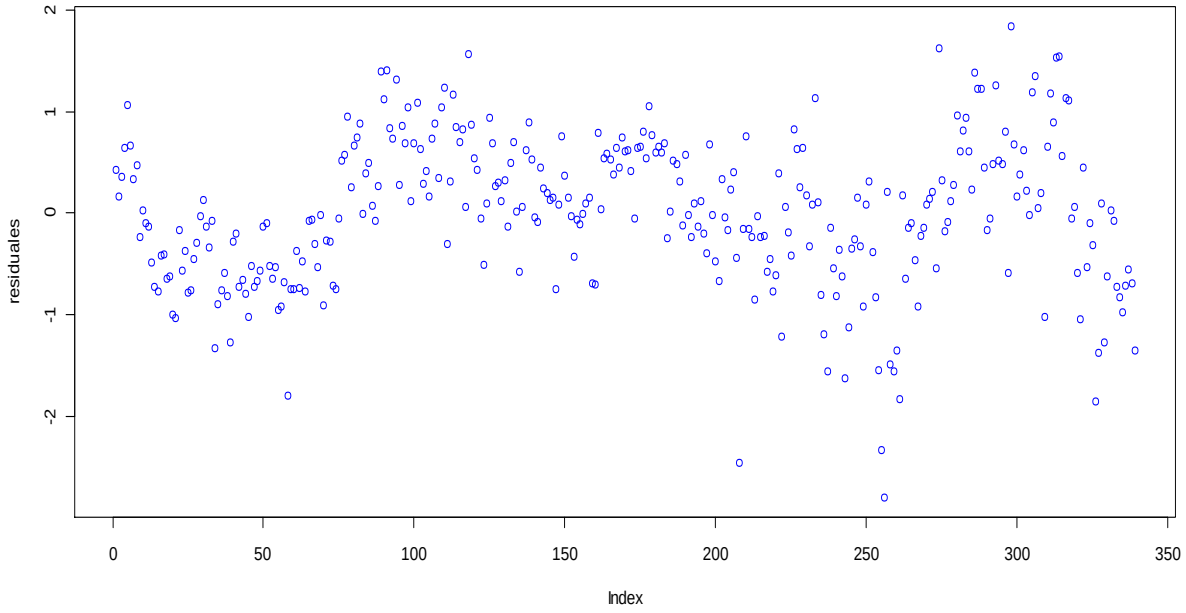
VARIABLE	FACTOR INFLACIÓN DE VARIANZA (VIF)
Empleo (emp)	***45.629***
Índice Manufactura (imf)	***99.739***
Índice Comercio Exterior (ice)	5.406
Índice de Turismo (itu)	***93.907***
Registro vehículos motor (vmotor)	2.311
Ventas al detalle (vad)	***10.428***
Producción energía eléctrica (pee)	***16.513***
Índice de construcción (iconst)	***26.588***

Seis de las variables tienen VIF mayores de 10, por lo que exhiben problemas de multicolinealidad. Nótese que algunos de estos valores VIF son particularmente altos. Las únicas variables que no muestran esta problema son la del índice de comercio exterior (Ice) y registro de vehículos de motor (Vmotor).

Para verificar la presencia de autocorrelación, se utilizaron dos procedimientos: a) producir la gráfica los residuales vs. los valores del IAE y b) por la prueba de Durbin-Watson.

FIGURA 4

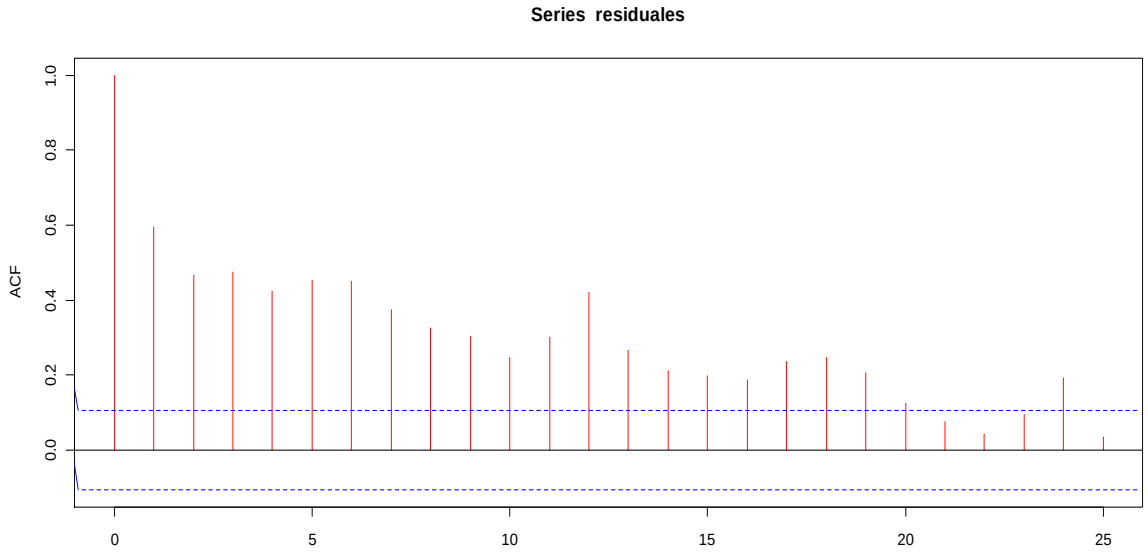
Gráfica de Residuales



De la gráfica de residuales (Figura 4) se podría detectar la presencia del problema de autocorrelación, porque los residuales tienden a formar ciclos.

FIGURA 5

Gráfica del Regresograma



Con la Figura 5 se confirma que existe un problema de autocorrelación.

El segundo procedimiento usado es una forma matemática de corroborar la presencia de autocorrelación utilizando la Prueba de Durbin-Watson.

- **Prueba de Durbin-Watson**

  Estadístico D-W

  0.7981223

El valor del estadístico D-W de la Prueba de Durbin-Watson se aproxima a cero, lo que nos indica que hay problemas de autocorrelación.
Las conclusiones que podemos sacar después de analizar multicolinealidad y autocorrelación son:

- El Índice de Actividad Económica desarrollado por La Junta de Planificación de Puerto Rico muestra serios problemas porque las variables regresoras están altamente correlacionadas (problema de multicolinealidad) y el modelo presenta también autocorrelación.
- Existe un serio problema con la variable *empleo,* por que se repite en el cálculo de los índices de manufactura, turismo y construcción, ésta puede ser una las razones por lo que hay problemas de multicolinealidad.

Bajo las conclusiones obtenidas proponemos otra metodología de obtener el Índice de Actividad Económica de Puerto Rico, remplazando la variable empleo por la variable *Tasa de desempleo (tdp);* y en vez de utilizar los índices de manufactura, comercio exterior, turismo, construcción, se utilizarán algunas de las variables con las cuales se calcularon estos índices.

El Índice Actividad Económica propuesto (IAEP) estaría en función de la tasa de desempleo, nómina en la manufactura, exportaciones, importaciones, registro hotelero, venta de cemento, valor permiso de construcciones, ventas al detalle, producción de energía eléctrica, y registro de vehículos de motor nuevos .

$$IAEP = f(tdep, nommfg, expor, import, reghot, vetcemen, valperconst, Vad, Pee, Vmotor)$$

A este modelo se le hará un estudio exhaustivo de problemas de multicolinealidad y autocorrelación, los pasos a seguir son:

- Encontrar el número condición de la matriz, y, los VIF's para cada variable, para ver si hay multicolinealidad.
- En caso de encontrar multicolinealidad, hacer la limpieza de la información, eliminando la variable con el valor de VIF más alto, luego volver a calcular el número condición de la matriz y los VIF's, se realiza este proceso hasta encontrar valores pequeños para el número condición y los VIF's.

TABLA 4.3

Matriz de Correlaciones del modelo propuesto

	tdep	*import*	*reghot*	*vmotor*	*vetcemen*	*valperconst*	*nommf*	*expor*	*vad*	*pee*
tdep	1									
import	-0.785	1								
reghot	-0.800	**0.901**	1							
vmotor	-0.565	0.741	0.664	1						
vetcemen	-0.843	0.812	0.840	0.644	1					
valperconst	-0.773	**0.863**	0.820	0.647	0.819	1				
nommf	-0.859	**0.865**	0.858	0.649	0.850	0.838	1			
expor	-0.787	**0.967**	**0.903**	0.708	0.822	0.885	0.878	1		
Vad	-0.794	**0.958**	0.897	0.727	0.752	0.839	0.877	**0.952**	1	
Pee	-0.832	**0.932**	**0.901**	0.723	0.879	0.875	**0.919**	**0.939**	**0.924**	1

De la Tabla 4.3 podemos decir que hay algunas variables que tiene correlación alta,

Corroboraremos esto con el cálculo de numero de condición y los VIF's.

- **Numero condición**

  NC= 18.14788

  Este número no es muy alto pero aún así se puede afirmar que hay problemas de multicolinealidad en los datos.

- **Factor de inflación de varianza para cada una de las variables**

TABLA 4.4

Factores de Inflación de la Varianza (VIF) para el modelo propuesto

VARIABLE	FACTOR INFLACIÓN DE VARIANZA (VIF)
Tasa de Desempleo (tdep)	5.319
Importaciones (import)	23.037
Registro Hotelero (reghot)	7.752
Vehículo de Motor (vmotor)	2.464
Venta de Cemento (vetcemen)	9.458
Valor de permisos de contrucciones (valperconst)	5.405
Nómina de manufactura (nommf)	8.467
Exportaciones (expor)	23.339
Ventas al Detalle	24.766
Producción de Energía Eléctrica	17.445

De los valores del factor de inflación de varianza (TABLA 4.4) se puede afirmar que las variables que causan multicolinealidad son ventas al detalle (*vad*), exportación (*expor*), importaciones (*import*) y Producción de energía eléctrica (*pee*).

Una vez realizado el ajuste en el modelo, las variables que quedan son: tasa de desempleo (*tdep*), importaciones (*import*), registro hotelero (*reghot*), vehículo de motor (*vmotor*), venta de cemento (*vetcemen*), valor de los permisos de construcción (*valperconst*) y nómina de manufactura (*nommf*).

TABLA 4.5

Matriz de Correlaciones del modelo final

	tdep	*import*	*reghot*	*vmotor*	*vetcemen*	*valperconst*	*nommf*
tdep	1						
import	-0.785	1					
reghot	-0.800	**0.901**	1				
vmotor	-0.565	0.741	0.664	1			
vetcemen	-0.843	0.812	0.840	0.644	1		
valperconst	-0.773	0.863	0.820	0.647	0.819	1	
nommf	-0.859	0.865	0.858	0.649	0.850	0.838	1

De la Tabla 4.3 podemos afirmar que las variables no están poco correlacionadas, para corroborar esto calculamos el número condición de la matriz y los respectivos VIF's.

- **Número de condición de la matriz**

  NC= 8.646427

  Este valor es menor que 10, que como ya se indicó es la cota superior de los NC para considerar que no hay problema de multicolinealidad en un conjunto de variables.

- **Factor de inflación de varianza para cada una de las variables**

TABLA 4.6

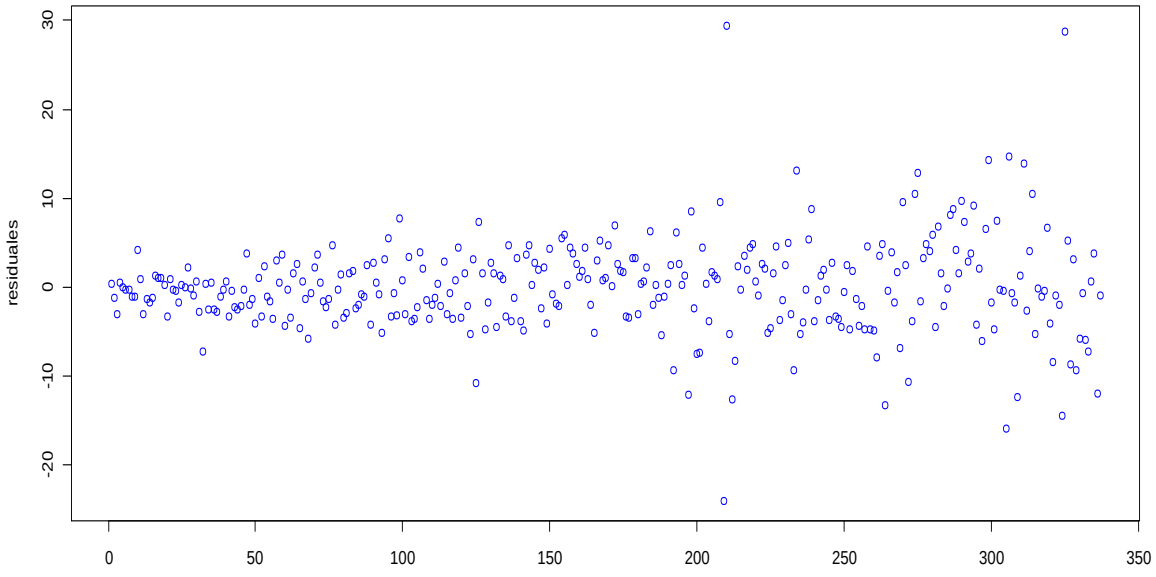
Factores de Inflación de la Varianza (VIF) modelo final
Variables del IAE

VARIABLE	FACTOR INFLACIÓN DE VARIANZA (VIF)
Tasa de Desempleo (tdep)	4.718
Importaciones (import)	8.777
Registro Hotelero	6.807
Registro Vehículo de Motor (vmotor)	2.287
Venta de Cemento (vetcemen)	5.421
Valor de Permisos de Construcción (valperconst)	4.864
Nómina de Manufactura (nommf)	6.751

Los VIF's para las variables del modelo final (TABLA 4.6) son bastante pequeños (menores que 10). Esto corrobora que no hay presencia de multicolinealidad.

FIGURA 6

Gráfica de Residuales Modelo Propuesto



Del gráfico de residuales (FIGURA 6) se puede afirmar que no hay problemas de autocorrelación en los residuales; sin embargo, pareciera que hay problema de varianzas no constantes en los residuales , para verificar esto se calcula la prueba de Breusch-Pagan para las hipótesis

H_o: $Las\ varianzas\ de\ los\ residuales\ son\ iguales\ (Homocedasticidad)$
H_a: $Por\ lo\ menos\ una\ de\ las\ varianzas\ difiere\ de\ las\ demás\ (Heterocedasticidad)$

Prueba de Breusch-Pagan

studentized Breusch-Pagan test

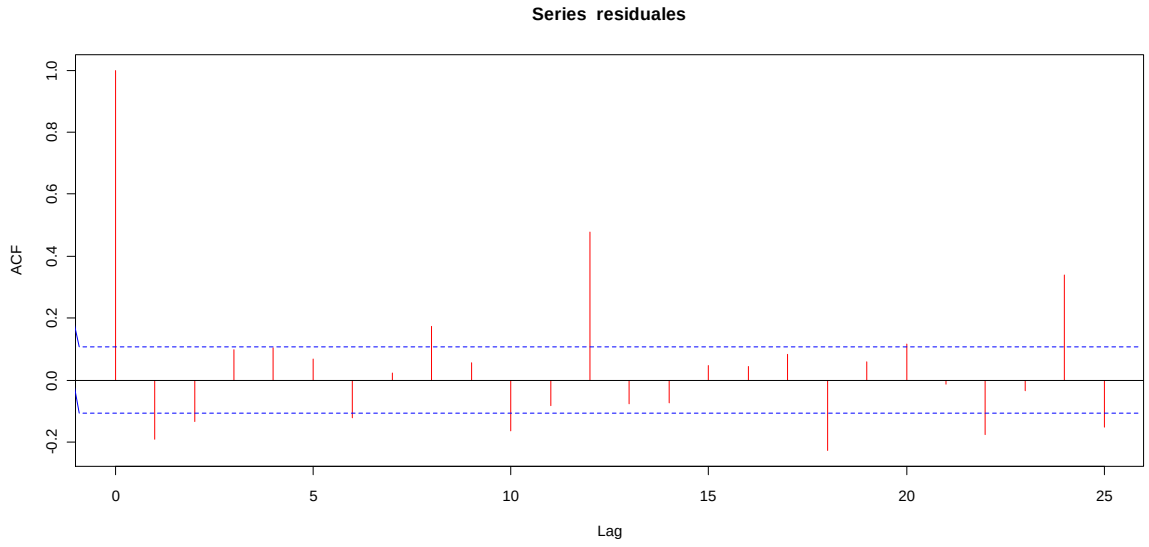
data: lm(l1)

BP = 23.5634, df = 7, p-value = 0.042329

Para un $\alpha = 0.05$ la prueba de Breusch-Pagan indica que a ese nivel significancia hay por lo menos una varianza de los residuales que difiere de las otras. Como el P-value para la prueba Breusch-Pagan no es muy pequeño con respecto al nivel de significancia el problema de heterocedasticidad pudiera no ser tomado en cuenta.

FIGURA 7

Gráfica del Regresograma Modelo Propuesto



La gráfica del regresograma (FIGURA 7) nos ayuda a interpretar que no hay presencia de autocorrelación.

Prueba de Durbin-Watson

D-W Statistic

2.3547

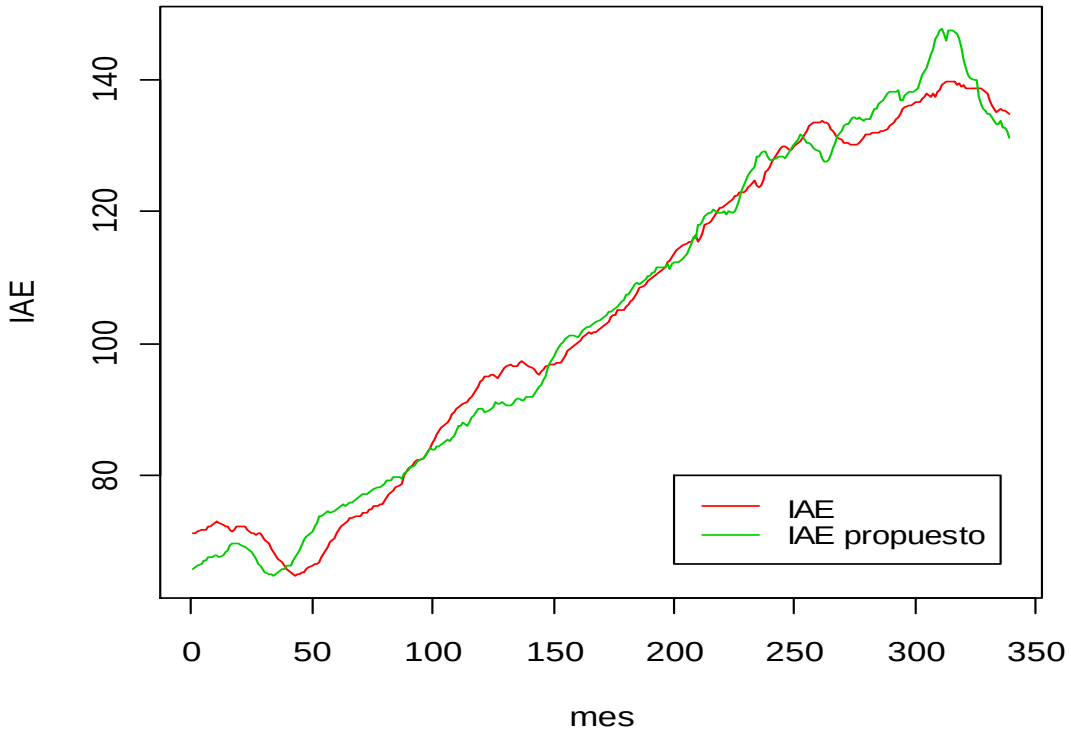
La prueba de Durbin-Watson resulta ser un valor cercano a 2, eso nos ratifica que no hay autocorrelación en los errores.

Dado que se solucionó el problema de multicolinealidad y autocorrelación, entonces el Índice de Actividad Económica que se propone para Puerto Rico estará determinado por la siguiente función.

$$IAEpropuesto = f(tdep, import, reghot, Vmotor, vetcemen, valperconst, nommfg)$$

FIGURA 8

Gráfica del IAE propuesto VS IAE



La Figura 8 es una comparación del Índice de Actividad Económica de Puerto Rico con la que trabaja la Junta de Planificación Económica y el Índice de Actividad Económica propuesto en el presente trabajo de investigación.

Regresión Logística para el IAEP

En esta sección se utilizará la técnica Estadística Multivariante de Regresión Logística para el modelo final del cálculo del Índice de Actividad Económica de Puerto Rico. Donde la variable respuesta es una variable dicotómica toma el valor de cero cuando se trata de un mes en recesión económica y tomará el valor de uno cuando se trate de un mes de expansión económica.

$$y = f(tdep, import, reghot, Vmotor, vetcemen, valperconst, nommfg)$$

TABLA 4.7

Coeficientes de la función logística

Constante	2.6624972
Tasa de desempleo(tdep)	-0.0920538
Importaciones(import)	0.0007170
Registro hotelero(reghot)	-0.0067658
Vehículo de motor(vmotor)	0.4032768
Venta de cemento(vetcemen)	-1.3123789
Valor de permisos de construcción (valpco)	-0.0009957
Nómina de manufactura(nommf)	-0.0046504

La Tabla 4.7 representa los coeficientes de la Regresión Logística

La función logística queda como:

$$\pi_i = \frac{e^{2.66-\text{tdep}0.09+\text{import}0.00072-\text{reghot}0.007+\text{vmotor}0.403-\text{vetcemen}1.31-\text{valpco}0.0010-\text{nommf}0.005}}{1+e^{2.66-\text{tdep}0.09+\text{import}0.00072-\text{reghot}0.007+\text{vmotor}0.403-\text{vetcemen}1.31-\text{valpco}0.0010-\text{nommf}0.005}}$$

Degrees of Freedom: 355 Total (i.e. Null); 348 Residual

Null Deviance: 352.9

Residual Deviance: 238.4 AIC: 254.4

Verificamos la confiabilidad del modelo de Regresión Logística para las hipótesis

H_o: $El\ modelo\ ajustado\ es\ confiable$

H_a: $El\ modelo\ ajustado\ no\ es\ confiable$

$$\chi^2_{\alpha,n-p-1} = \chi^2_{0.05,348} = 392.5$$

Donde D=238.4, como el valor de la Devianza es menor que $\chi^2_{0.05,348} = 392.5$, entonces se puede concluir que el modelo de Regresión Logística es un modelo confiable para la base de datos.

$$Pseudo - R^2 = 1 - \frac{Devianza\ Residual}{Devianza\ Nula} = 1 - \frac{238.4}{352.9} = 0.324$$

Como el Pseudo-R^2 es mayor que 0.3, entonces el modelo de Regresión Logística es un modelo aceptable para la base de datos.

FIGURA 9

Gráfica de Sensitividad y Especificidad

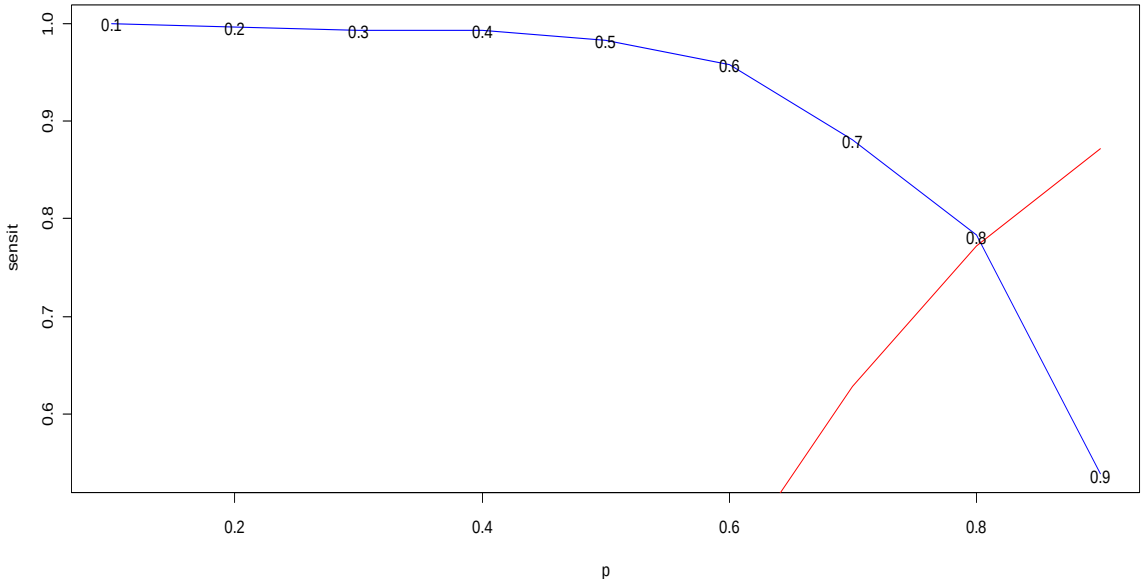
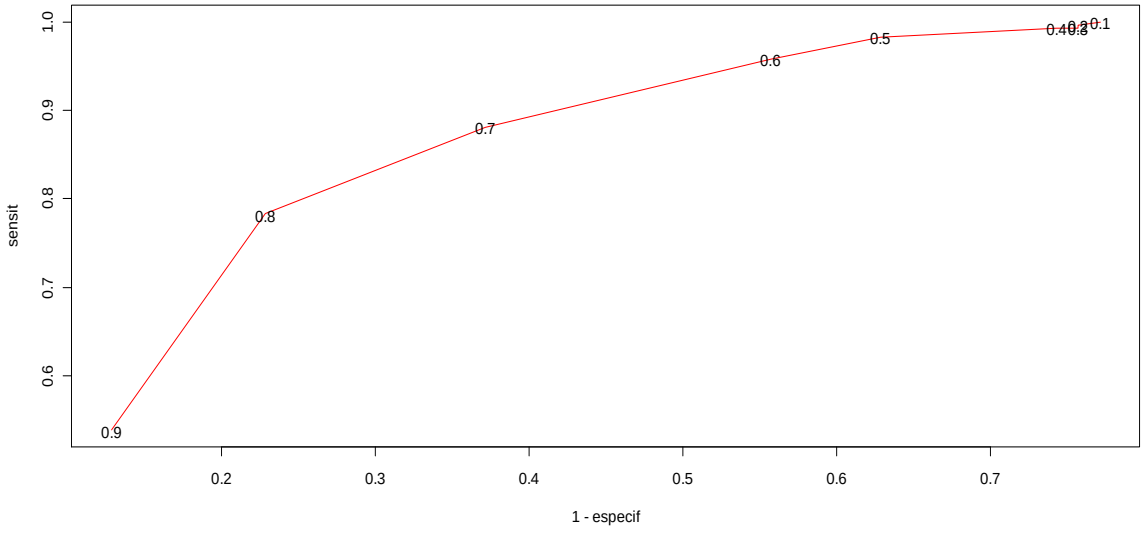


FIGURA 10

Gráfica del ROC



De las Figuras 9 y 10 se puede decir que el p óptimo es 0.8, entonces todos los que tienen una probabilidad superior a 0.8 serán clasificados en la clase 1 de expansión y los otros en la clase 0 de recesión, bajo este criterio la *tasa de mala clasificación óptima es 0.2191*

Análisis discriminante para el IAEP

En esta sección se utilizará la técnica Estadística Multivariante de Análisis Discriminante para el modelo final del cálculo del Índice de Actividad Económica de Puerto Rico. Donde la variable respuesta es una variable dicotómica toma el valor de cero cuando se trata de un mes en recesión económica y tomará el valor de uno cuando se trate de un mes de expansión económica.

$$y = f(tdep, import, reghot, Vmotor, vetcemen, valperconst, nommfg)$$

TABLA 4.8

Coeficientes de la función discriminante

Constante	-2.683
Tasa de desempleo(tdep)	0.1323919
Importaciones(import)	0.0000408
Registro hotelero(reghot)	0.0040027
Vehículo de motor(vmotor)	-0.1006958
Venta de cemento(vetcemen)	0.3567055
Valor de permisos de construcción (valperconst)	0.0002309
Nómina de manufactura(nommf)	0.0027663

La Tabla 4.8 representa a los coeficientes de la función discriminante.

La función discriminante es:

$$Y = -2.683 + tdep * 0.1323919 + import * 0.0000408 + reghot * 0.0040027$$
$$-vmotor * 0.100696 + vetcemen * 0.356706 + valperconst * 0.000231 + nommf * 0.0028$$

TABLA 4.9

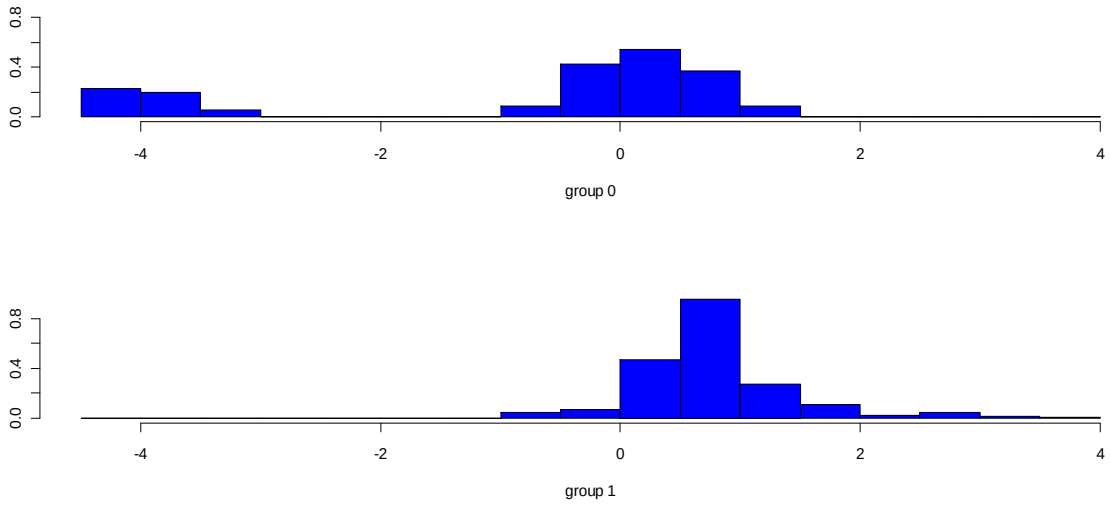
Probabilidades a priori de los grupos

0	1
$P(G_1 = 0)$ =0.196629	$P(G_2 = 1)$=0.803370

La Tabla 4.9 representa a las probabilidades a priori, donde $P(G_1 = 0)$ es la probabilidad de pertenecer al grupo de recesión y $P(G_2 = 1)$ es la probabilidad de pertenecer al grupo de expansión.

FIGURA 11

Gráfica de las Puntuaciones Discriminantes



La gráfica de las puntuaciones discriminantes (FIGURA 11) por grupos, indica que en muchos casos las puntuaciones van a separar correctamente los meses de recesión y expansión. Hay una zona en la que los dos grupos se solapan, esta es la zona en que se incrementa el valor de la tasa de mala clasificación.

TABLA 4.10

Coeficientes estandarizados

Tasa de desempleo	**0.514**
Importaciones	-0.041
Registro Hotelero	0.164
Vehículo de motor	**-0.591**
Venta de cemento	0.259
Valor de permiso de construcción	0.019
Nómina de manufactura	**0.898**

La Tabla 4.10 lleva a establecer la relación entre las variables explicativas y la función discriminante, con el fin de determinar qué variables son las que aportan más a la discriminación a través de los coeficientes estandarizados, ignorando su signo, todos aquellos coeficientes que tengan valores altos serán los que contribuyen de manera significativamente al momento de hacer la discriminación. Los que mayor valor estandarizado tienen se muestran en negrita en la tabla.

TABLA 4.11

Matriz de estructura

Tasa de desempleo	0.159
Importaciones	0.019
Registro Hotelero	-0.046
Vehículo de motor	**-0.537**
Venta de cemento	-0.260
Valor de permiso de construcción	-0.114
Nómina de manufactura	**0.756**

La Tabla 4.11 representa la correlación lineal entre cada variable original y la función discriminante, aquellas variables que tienen coeficientes de correlación alta son las que contribuyen significativamente en la discriminación.

De las Tablas 4.10 y 4.11 se puede concluir que las variables Vehículos de Motor y Nómina de Manufactura, son las que contribuyen de manera significativa en la discriminación de grupos.

TABLA 4.12

Clasificación de resultados por error aparente

	Asignados por el Modelo de A.D.			
Asignados por la Junta de Planificación de Puerto Rico		0 Recesión	1 Expansión	Totales
	0 Recesión	35	35	70(20%)
	1 Expansión	16	270	286(80%)
	Totales	51(14%)	305(86%)	356

La Tabla 4.12 muestra el número de datos que fueron mal clasificados, en el grupo 0 (recesión) 35 meses que hubo recesión fueron clasificadas como meses de expansión económica. En el grupo 1 (expansión) 16 meses de expansión económica fueron clasificados como meses de recesión. Por tanto la tasa de mala clasificación es 0.143258.

TABLA 4.13

Clasificación de resultados por Validación Cruzada

	Asignados por el Modelo de A.D.			
Asignados por la Junta de Planificación de Puerto Rico		0 Recesión	1 Expansión	Totales
	0 Recesión	34	36	70(20%)
	1 Expansión	17	269	286(80%)
	Totales	51(14%)	305(86%)	356

La Tabla 4.13 muestra el número de datos que fueron mal clasificados por el método de validación cruzada, en el grupo 0 (recesión) 36 meses que hubo recesión fueron clasificadas como meses de expansión económica. En el grupo 1 (expansión) 17 meses de expansión económica fueron

clasificados como meses de recesión económica. Por tanto la tasa de mala clasificación es 0.148876.

Relación de las variables regresoras en periodos de expansión y periodos de recesión

En esta sección se hará un análisis de las variables regresoras en periodos de expansión y periodos de recesión.

TABLA 4.14

Matriz de correlación para periodos de expansión.

	tdesempleo	import	reghot	vmotor	vetcemen	valperconst	nommf
tdesempleo	1.000						
import	-0.770	1.000					
reghot	-0.791	**0.896**	1.000				
vmotor	-0.514	0.718	0.644	1.000			
vetcemen	**-0.826**	0.799	**0.830**	0.618	1.000		
valperconst	-0.759	**0.849**	**0.802**	0.615	**0.808**	1.000	
nommf	**-0.851**	**0.852**	**0.852**	0.608	**0.847**	**0.821**	1.000

En la Tabla 4.14 se puede observar las correlaciones de las variables regresoras en periodos de expansión.

TABLA 4.15

Matriz de correlación para periodos de recesión.

	tdesempleo	import	reghot	vmotor	vetcemen	valperconst	nommf
tdesempleo	1.000						
import	**-0.837**	1.000					
reghot	**-0.838**	**0.886**	1.000				
vmotor	-0.651	0.413	0.520	1.000			
vetcemen	**-0.803**	0.598	0.712	0.761	1.000		
valperconst	**-0.827**	**0.860**	**0.832**	0.648	0.718	1.000	
nommf	-0.602	**0.886**	0.716	0.098	0.234	0.645	1.000

En la Tabla 4.15 se puede observar la correlación de las variables regresoras en periodos de recesión.

5.- CONCLUSIONES

Al finalizar el presente trabajo de investigación se llegó a las siguientes conclusiones.

El modelo del IAE que calcula la Junta de Planificación Económica de Puerto Rico muestra algunos problemas: heterocedasticidad de los errores; autocorrelación y de multicolinealidad entre las variables explicativas, pues éstas están altamente correlacionadas. Sin embargo el modelo que se plantea en este trabajo resulta ser mejor desde el punto de vista estadístico por que las variables no están altamente correlacionadas y además no existe problema de autocorrelación, aunque sí se presenta un ligero problema de heterocedasticidad.

Haciendo uso de las técnicas estadísticas multivariantes, según el Análisis Discriminante, las variables vehículo de motor y nómina de manufactura son las que contribuyen de manera significativa en la discriminación de grupos. En la Regresión Logística las variables vehículo de motor, venta de cemento y nómina de manufactura, son las que contribuyen de manera significativa en la discriminación.
Al comparar estas dos técnicas estadísticas multivariantes el Análisis Discriminante arroja una menor tasa de mala clasificación de 15% contra 22%.

Para el modelo que se propone en este trabajo de investigación se puede afirmar que en periodos de recesión la variable tasa de desempleo está altamente correlacionada y con signo negativo con el resto de las variables con la excepción de vehículos de motor; así como también la variable Importaciones está altamente correlacionada de manera positiva con las variables registro hotelero, valor de permisos de construcción y nómina en manufactura (Tabla 4.15).
Para los periodos de expansión la variable nómina en manufactura se encuentra altamente correlacionada positivamente con las variables importaciones, registro hotelero, venta de cemento y valor de los permisos de construcción, además está correlacionada negativamente con la variable tasa de desempleo. La variable valor de permisos de construcción se encuentra altamente correlacionada positivamente con Importaciones, registro hotelero y venta de cemento (Tabla 4.14).

6.- TRABAJO FUTURO

Como trabajos futuros planteamos los siguientes:

La verificación matemática al momento de comparar dos ciclos económicos.

Utilizar otras técnicas multivariadas como Análisis Factorial para verificar cuáles son los factores que influyen más en el estudio de los Ciclos Económicos de Puerto Rico.

Utilizar la Regularización de Tikhonov para los solucionar los problemas de multicolinealidad.

7.- BIBLIOGRAFÍA

[1] Acuña, E. (2009), Análisis de Regresión. Departamento de Matemáticas, Universidad de Puerto Rico-Recinto Universitario de Mayagüez.

[2] Hosmer, D y Lemeshow, S. (2000), Applied Logistic Regression, Second Edition.

[3] Montañés, A. (1995), Econometría I, Facultad de Ciencias Económicas y Empresariales de la Universidad de Zaragoza.

[4] Cuevas, A. y Berrendero, J., (2003), Análisis Discriminante: Prácticas con R. Departamento de Matemáticas Universidad Autónoma de Madrid.

[5] Wschebor, M. y Armentano, D., (2005), Número de Condición y Matrices Aleatorias, Facultad de Ciencias Universidad de la República de Montevideo.

[7]Belsley, D., Kuh, y Welsh, R., (1980), Regression Diagnostics. John Wiley, New York.

[8]Draper, N. y Smith, H., (1998), Applied Regression Analysis, Third Edition. John Wiley, New York.

[9]Mardia, K., Kent, J. y Bibby, J., (1980), Multivariate Analysis.

[10]Johnson, R. y Wichern, D. Applied Multivariate Statistical Analysis.

[11]Tusell, F., (2008), Análisis Multivariante, First Edition.

[12]Klijn, F.,(2001), Análisis Multivariante.

[13]Brockwell, P.J. y Davis, R.A.,(2002), Introduction to Times Series and Forecasting, Second Edition.

[14] Alameda, J., (2008), Ensayos en Economía Aplicada.

[15]Bram, J. y Martínez, F. E. y Steindel, Ch., (2008), Tendencias y Cambios en la Economía de Puerto Rico.

[16] Curet, E.,(1976), El Desarrollo Económico de Puerto Rico

[17] Alameda, J.I.,(2005), Un Modelo de Estimación de la Probabilidad de una Recesión para Puerto Rico

[18] Alameda, J.I., la Recesión 2005-2007, Departamento de Economía, Recinto Universitario de Mayagüez

[19] Alameda, J.I., Los Ciclos Económicos en Puerto Rico: Cronología y Medición.

8.- ANEJOS

Anejo 1 Regresión Logística.

```
read(data)
logis<-glm(y~tdesempleo+import+reghot+vmotor+vetcemen+valperconst+nommf
,data=data,family =binomial)
summary(logis)
```

```
Call:
glm(formula = y ~ tdesempleo + import + reghot + vmotor + vetcemen +
   valperconst + nommf, family = binomial, data = iriss)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.67557   0.05295   0.32937   0.58969   1.85034

Coefficients:
                 Estimate    Std. Error    z value   Pr(>|z|)
(Intercept)    2.6624972   2.7108414    0.982    0.32602
tdesempleo  -0.0920538  0.0899705   -1.023    0.30623
import          0.0007170  0.0007783    0.921    0.35693
reghot         -0.0067658  0.0142269   -0.476    0.63439
vmotor          0.4032768  0.0795627    5.069    4.01e-07 ***
vetcemen    -1.3123789   0.6264592   -2.095    0.03618 *
valperconst  -0.0009957   0.0056587   -0.176    0.86033
nommf        -0.0046504  0.0015758   -2.951    0.00317 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 352.93  on 355  degrees of freedom
Residual deviance: 238.39  on 348  degrees of freedom
AIC: 254.39

Number of Fisher Scoring iterations: 7
```

Haciendo la clasificación con el método mas complicado calculando la sensitividad y especificidad

```
p<-seq(.1:.9,length=9,by=.1)
sensit<-rep(0,9)
especif<-rep(0,9)
```

```
for(j in 1:9)
{
            clases1<-rep(0,nobs)
             for(i in 1:nobs)
            {
             if(phat[i]>=p[j]){clases1[i]<-1
              }
 }
data1<-cbind(data[,1],clases1)
exp<-data1[data1[,1]==1,]
rece<-data1[data1[,1]==0,]
sensit[j]<-mean(exp[,1]==exp[,2])
especif[j]<-mean(rece[,1]==rece[,2])
}
tabla<-cbind(p,sensit,especif)
cat("Sensitividad y especifidad p\n")
print(tabla)
```

Plot de sentividad y especificidad para encontrar el p óptimo.

```
win.graph()
plot(p,sensit,type="l",col="blue")
lines(p,especif,col="red")
text(p,sensit,labels=p,col="9")
title("Gráfica de sentividad y especificidad")
```

Plot de la curva ROC

```
win.graph()
plot(1-especif,sensit,type="l",col="red")
text(1-especif,sensit,labels=p,col="9")
```

```
title("Curva ROC")
```

Clasificación utilizando el p óptimo.

```
clasesf<-rep(0,nobs)
for(i in 1:nobs)
{
  {
    if(phat[i]>=0.8){clasesf[i]<-1
    }
}
erroresf<-sum(clasesf!=data[,1])
ratef<-erroresf/nobs
cat("tasa de mala clasificacion",ratef,"\n")
```

Anejo 2 Análisis Discriminante

```
read(data)
library(MASS)   # Cargamos la libreria que contiene a  lda
data(data)      # Cargamos los datos
data
dim(data)
datos = data.frame(iriss[,1:7],clase=as.vector(iriss[,8]))
lda(clase~.,datos)
data.lda = lda(clase~.,datos)
plot(data.lda, col="blue")
attributes(data.lda)      # cosas que devuelve la función
data.lda$prior
```

Tasa de mala clasificación por Error Aparente

```
table(data[,8],predict(data.lda,data[,1:7])$class)
predclas<-predict(data.lda)$class
tea<-sum(predclas==clases)/356
cat("tasa de mala clasificacion",ratef,"\n")
```

Tasa de mala clasificación por Validación Cruzada

```
library(MASS)
predclas.cv<-lda(data,clases,CV=T)$class
tevc<-1-sum(predclas.cv==clases)/356
tevc
```

Anejo 3 Prueba de Durbin-Watson

```
red(data)
l1<-lm(IAEpropuesto~tdesempleo+import+reghot+vmotor+vetcemen+valperconst+nommf
,data=data)
residuales=l1$res
au=acf(residuales, plot=FALSE)
plot(au,col="red")
title(“ Series Residuales”)
plot(residuales, col="blue")
title("Gráfica de residuales")
dw<-function(durbin)
{
    durbin1<-durbin[-1]
    durbin2<-durbin[-length(durbin)]
    diff<-durbin1-durbin2
    dw<-sum(diff^2)/sum(durbin^2)
    dw
```

```
}
DW=dw(residuales)
```

Anejo 4 Numero condición y Factor de Inflación de Varianza

```
read(data)
attach(data)
matrizcor<-cor(data[,1:n])
ev=eigen(cor(data[,1:n]))
eigvals<-ev$values
eigvals
cond=sqrt(eigvals[1]/ev$values[n])
cond
vif<-diag(solve(matrizcor))
cat("los VIF's son:\n")
vif
```

Anejo 5 Gráfico comparativo del IAE

```
read(data)
ts.plot(data,col=2:3,xlab="mes", ylab="IAE", main="Índice de Actividad Económica")
bandas <- expression("ICAE", "ICAE propuesto")
legend(200,80, bandas,lty=1, col=c(2,3),cex=.9)
```