INTRODUCCION AL ANÁLISIS DE REGRESIÓN LINEAL

Larissa Welti Santos

Cholula, México. 2002

El análisis de regresión tiene por objetivo estimar el valor promedio de una variable, variable dependiente, con base en los valores de una o más variables adicionales, variables explicativas. En este tipo de análisis, la variable dependiente es estocástica mientras que las variables explicativas son no estocásticas en su mayor parte[1]. El análisis de regresión ha cobrado popularidad debido al gran número de paquetes estadísticos que lo incluyen y por ser un "proceso robusto que se adapta a un sinfín de aplicaciones científicas y ejecutivas que permite la toma de decisiones" (Linne et al. 2000, p. 47, tr.). En este trabajo, el mejor ajuste de los modelos estará determinado por el análisis de regresión lineal.

1.1 Modelo de regresión

Considérese la siguiente relación para explicar el comportamiento de una variable dependiente (*Y*) en función de *n* variables dependientes $(X_1, X_2, \ldots, X_n)$.

$$Y = f(X_1, X_2, \ldots, X_n) \tag{1.1.1}$$

donde $f(\ \ )$ es una forma funcional implícita. En el caso en el cual esta forma funcional no pueda estimarse, $f(\ \ )$ puede aproximarse mediante:

$$Y = \sum_{i=1}^{n} \beta_{i+1} X_i + \Psi \tag{1.1.2}$$

[1] "En el tratamiento avanzado se puede liberar el supuesto de que las variables explicativas no son estocásticas" (Gujarati,1990).

para $i = 1,2,\ldots,n$ donde las β son parámetros de la función y Ψ es el error debido a la aproximación lineal de (1.1.1).

En la realidad, la ecuación (1.1.2) no se cumple estrictamente pues existe también una variación en la variable dependiente debido a que hay errores de medición. A esta variación inexplicable se le denomina ruido blanco y se denota como η. Por otro lado, no todas las variables independientes son medibles o se puede tener acceso a la información por lo que sólo algunas de ellas se utilizarán finalmente en el modelo.

Supóngase se tiene una muestra de *m* observaciones, $j = 1,2,\ldots,m$ e información sobre *k* variables independientes que determinan en parte el comportamiento de *Y*. La ecuación (1.1.2) puede rescribirse como:

$$Y_j = \sum_{i=1}^{k} \beta_{i+1} X_{ij} + \Psi_j + \eta_j + \mathrm{Z}_j \tag{1.1.3}$$

donde Z es el efecto de las *n* - *k* variables ($k < n$) que no fueron incluidas en el modelo.

Sean $\Psi_t = \overline{\Psi} + \psi_t$ y $\mathrm{Z}_t = \overline{\mathrm{Z}} + z_t$ donde ψ y z son las desviaciones con respecto a las medias de Ψ y Z respectivamente, entonces:

$$Y_j = \beta_1 + \beta_2 X_{1j} + \beta_3 X_{2j} + \ldots + \beta_{k+1} X_{kj} + \varepsilon_j \tag{1.1.4}$$

donde

$$\beta_1 = \overline{\mathrm{Z}} + \overline{\Psi} \text{ y } \varepsilon_j = z_j + \psi_j + \eta_j .$$

La ecuación (1.1.4) se conoce como la ecuación de regresión lineal múltiple, donde las β son los coeficientes de la regresión que necesitan estimarse y las X las variables independientes.

1.1.1 Coeficientes de regresión

La ecuación (1.1.4) tiene k parámetros asociados a las variables independientes X. $\beta_2, \beta_3, \ldots, \beta_{k+1}$ se interpretan como las derivadas parciales de Y con respecto a las X i.e. $\partial Y / \partial X_i = \beta_i$. β_i dice qué tanto cambiará Y si ocurre un cambio unitario en X_i manteniendo todo lo demás constante. Sin embargo, los valores reales de estos coeficientes son desconocidos y habrá que estimarlos mediante algún método.

1.1.2 Constante de regresión

A diferencia de los otros coeficientes de la ecuación de regresión, β_1 no mide cambios, sino que corresponde al efecto medio en la variable dependiente Y que ocasionan tanto las variables que fueron excluidas en la ecuación como la aproximación lineal. A diferencia de un modelo matemático donde el término constante representa el intercepto con la ordenada, en un modelo econométrico, la interpretación de la constante de regresión, como ya se vio, es distinta. Sólo en algunas ocasiones, como en el caso de las funciones de costo donde existen costos fijos, esta constante sí puede interpretarse como el intercepto.

1.1.3 Estimación de los coeficientes

Hasta ahora se ha hecho referencia a la interpretación de los coeficientes pero no se ha hablado sobre el problema de la estimación. El objetivo del análisis de regresión será

buscar la mejor estimación de los parámetros para construir una aproximación cercana al Y real.

Supóngase que mediante algún procedimiento se obtuvieron las estimaciones de las β, ($\hat{\beta}$). El residual e_j se definirá como la diferencia entre el valor observado de Y_j y la predicción $\hat{Y}_j$ con base en los valores estimados de las β.

$$e_j = Y_j - \hat{\beta}_1 - \hat{\beta}_2 X_{1j} - \ldots - \hat{\beta}_{k+1} X_{kj} \tag{1.1.5}$$

donde $Y_j = \beta_1 + \beta_2 X_{1j} + \beta_3 X_{2j} + \ldots + \beta_{k+1} X_{kj} + \varepsilon_j$ es la ecuación de regresión múltiple.

Debido a que el residual e_j es una estimación de ε_j, éstos deben cumplir con las mismas propiedades de los errores ε_j:

1: $E\left[\varepsilon_j \middle| X_{ij}\right] = 0$.

2: $E\left[\varepsilon_i \varepsilon_j\right] = 0$. (no hay autocorrelación entre los errores)

3: $Var\left[\varepsilon_j \middle| X_{ij}\right] = \sigma^2$. (homocedasticidad)

4: $Cov\left[\varepsilon_j, X_{ij}\right] = 0$.

Ya que diferentes valores de $\hat{\beta}$ originarán residuales e_j distintos, se buscará el conjunto de $\hat{\beta}$ que produzcan residuales con las características anteriormente descritas, es decir: $\sum_{j=1}^{m} e_j = 0$; $\sum_{j=1}^{m} X_{ij} e_j = 0 \quad \forall i, i = 1,2,\ldots,k$.

Cuando la ecuación de regresión incluye el término β_1, el problema se reduce a minimizar la suma de cuadrados de los residuales, i.e. $\sum_{j=1}^{m} e_j^{\,2}$, y a este procedimiento se le conoce como Mínimos Cuadrados Ordinarios (MCO).

La estimación de las β utilizando MCO lleva consigo la restricción de que para obtener una solución no trivial el número de observaciones debe ser mayor al número de parámetros a estimar. La diferencia entre el número de observaciones y el número de parámetros a ser estimados se le denomina grados de libertad.

1.2 Estadísticas y pruebas de hipótesis

Una vez que se obtuvieron las estimaciones de los parámetros en una regresión lineal, se hace uso de estadísticas para juzgar la bondad del modelo, la utilidad de las estimaciones y la precisión de las mismas. Lo que a continuación se presenta son las herramientas básicas para el análisis de regresión.

1.2.1 Coeficiente de determinación

Si todas las observaciones coincidieran con la ecuación de regresión, se tendría un ajuste perfecto; sin embargo, rara vez sucede esto. Generalmente habrá e_i positivos y negativos por lo que se requiere de una medida que establezca qué tan bien la ecuación de regresión representa a los datos. El coeficiente de determinación R^2 es una medida de bondad de ajuste y se define de la siguiente manera:

$$R^2 = \frac{\sum_{j=1}^{m}(\hat{Y}_j - \overline{Y})^2}{\sum_{j=1}^{m}(Y_j - \overline{Y})^2} \qquad 0 \le R^2 \le 1 . \tag{1.2.1}$$

Donde el numerador corresponde a la suma de cuadrados debido a la regresión (SCE) y el denominador a la suma de cuadrados total (SCT). A medida que la SCE explique en gran parte la variación de Y_j, R^2 se acercará a uno. A pesar de que este coeficiente es una medida de bondad de ajuste no debe abusarse de él, pues R^2 puede aumentar agregando al modelo variables explicativas adicionales aunque no sean significativas.

1.2.2 Errores estándar

Dado que los estimadores de mínimos cuadrados están en función de la información muestral, es necesario encontrar la precisión de las $\hat{\beta}$. La manera convencional de medir la precisión de un estimador es por medio de su varianza. Entre más pequeña sea la varianza de un estimador mayor es su precisión, esto significa que los estimadores serán poco sensibles a los errores que pudieran existir en la muestra de la variable dependiente *Y*.

En forma teórica la varianza de β_i para una muestra de *m* observaciones y *k* variables independientes estará dada por la varianza de los errores divida por el elemento *i*-ésimo de la diagonal de la siguiente matriz donde $x_{ij} = X_{ij} - \overline{X_i}$.

$$\sigma_{\varepsilon}^{2}\begin{bmatrix} \sum_{j=1}^{m} x_{1j}^{2} & \sum_{j=1}^{m} x_{1j}x_{2j} & \cdots & \sum_{j=1}^{m} x_{1j}x_{kj} \\ \sum_{j=1}^{m} x_{2j}x_{1j} & \sum_{j=1}^{m} x_{2j}^{2} & \cdots & \sum_{j=1}^{m} x_{2j}x_{kj} \\ \vdots & \vdots & & \vdots \\ \sum_{j=1}^{m} x_{kj}x_{1j} & \sum_{j=1}^{m} x_{kj}x_{2j} & \cdots & \sum_{j=1}^{m} x_{kj}^{2} \end{bmatrix}^{-1}$$

Sin embargo, σ_{ε}^{2} es desconocida por lo que se estima mediante:

$$\hat{\sigma}_{\varepsilon}^{2} = \frac{\sum_{j=1}^{m} e_j}{m-k-1} \tag{1.2.2}$$

donde $m - k - 1$ son los grados de libertad.

Al sustituir (1.2.2) en la matriz anterior se obtiene una estimación de la varianza de $\hat{\beta}_i$. El error estándar de $\hat{\beta}_i$ se definirá como la raíz cuadrada de la estimación de la varianza de $\hat{\beta}_i$.

1.2.3 Significación de los coeficientes de regresión

No basta con saber qué tan bien se ajusta la línea de regresión a los datos ni con conocer los errores estándar de los parámetros estimados, es también muy importante conocer si la variable dependiente *Y* está realmente relacionada con la(s) *X*. Para ello se hace uso de pruebas de hipótesis donde se evalúa si los coeficientes relacionados a cada *X* son distintos de cero.

1.2.3.1 Modelo en dos variables

En el modelo de regresión lineal en dos variables se evalúa la siguiente hipótesis nula para saber si la variable *X* es o no significativa para la predicción de *Y*:

$H_0: \beta_2 = 0$ para un nivel de significación de α vs $H_a: \beta_2 \neq 0$.

La estadística de prueba es:

$$t_c = \frac{\hat{\beta}_2 - 0}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{j=1}^{m} x_j^{\,2}}}}. \qquad (1.2.3)$$

(1.2.3) se compara con una $t_{(m-2)}$ donde *m* es el número total de observaciones. Se rechaza H_0 si valor P < α[2]. En caso de rechazar H_0 se concluye que hay evidencia suficiente para afirmar que *X* está relacionada con *Y* a un nivel de significación de α.

Si se tienen expectativas previas del signo del coeficiente se establece la hipótesis alternativa como $H_a: \beta_2 < \beta_2^*$ ó $H_a: \beta_2 > \beta_2^*$. En estos casos, se rechaza H_0 si valor P < α[3].

1.2.3.2 Modelo con *k* variables

Se evalúa la siguiente hipótesis nula para saber si las variables independientes son significativas:

$H_0: \beta_2 = \beta_3 \ldots = \beta_{k+1} = 0$ para un nivel de significación de α vs H_a: al menos alguna β es distinta de cero.

---

[2] Valor P = $2\,P\left[t_{(m-2)} \geq |t_c|\right]$.

[3] Valor P = $P\left[t_{(m-2)} \leq t_c\right]$ ó Valor P = $P\left[t_{(m-2)} \geq t_c\right]$ según la hipótesis alternativa planteada.

La estadística de prueba es:

$$F = \frac{CME}{CMR} \tag{1.2.4}$$

donde *CME* es el cuadrado medio debido a la regresión y *CMR* es el cuadrado medio residual[4]. El resultado se compara con una $F_{k,m-k-1}$, (*k* son el número de variables independientes y *m* el número de observaciones) con un nivel de significación de α para poder decidir si se rechaza o no la hipótesis nula $\mathrm{H}_0 : \beta_2 = \beta_3 \ldots = \beta_{k+1} = 0$.

En el caso de la regresión múltiple no basta con probar que todos los coeficientes son significativamente distintos de cero, es necesario saber si agregar una variable al modelo una vez que existen otras incluidas no mejora significativamente la predicción de la variable dependiente. Para este caso se realizan pruebas parciales *F*.

Para probar H_0 : agregar X^* al modelo que ya tiene $X_1, X_2, \ldots, X_p$ variables independientes no mejora significativamente la predicción de la variable dependiente con un nivel de significación α, la estadística de prueba es:

$$F_c(X^*, X_1, X_2, \ldots, X_p) = \frac{SCE(X^*, X_1, X_2, \ldots, X_p) - SCE(X_1, X_2, \ldots, X_p)}{CMR(X^*, X_1, X_2, \ldots, X_p)} \tag{1.2.5}$$

[4] El cuadrado medio es igual a la suma de cuadrados dividida por los grados de libertad.

(1.2.4) se compara con una $F_{(1,m-p-2)}$ donde *m* es el número total de observaciones. Se rechaza H_0 si valor P < α[5] y se concluye que agregar X^* al modelo que ya tiene $X_1, X_2, \ldots, X_p$ variables independientes contribuye significativamente a la predicción de la variable dependiente[6].

[5] Valor P = $2\,P\left[F_{(m-p-2)} \geq |F_c|\right]$.

[6] La información de este capítulo se obtuvo de las publicaciones de Rao, M y Miller, R (1971) y Gujarati, D (1990).

ANEXOS

Anexo A Prueba *d* de Durbin-Watson

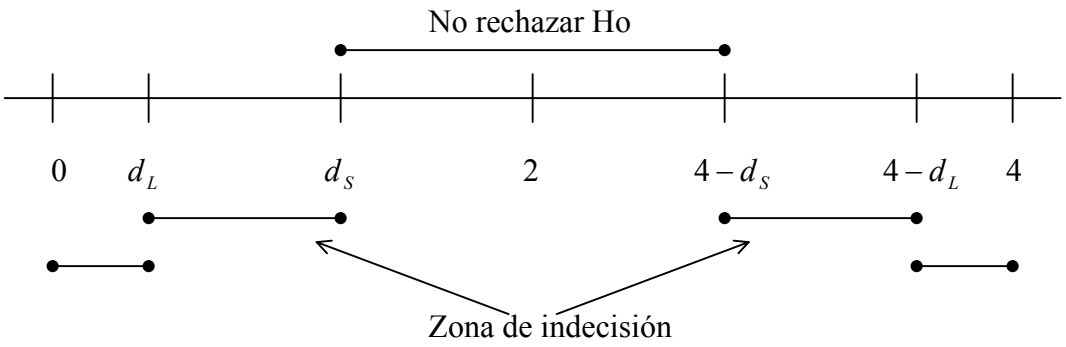
La prueba *d* de Durbin-Watson es la prueba más conocida para detectar la autocorrelación. El estadístico *d* de Durbin-Watson para *m* observaciones se define como:

$$d = \frac{\sum_{j=2}^{m} (e_j - e_{j-1})^2}{\sum_{j=1}^{m} e_j^{\,2}}. \tag{A.1}$$

Los supuestos en los que se basa este estadístico son:

1. El modelo de regresión incluye el término β_1.
2. Las variables explicativas son no estocásticas.
3. Los términos de error estocástico ε_j se generan a través de un esquema autorregresivo de primer orden, i.e. $\varepsilon_j = \rho\varepsilon_{j-1} + u_j$, $\mid \rho \mid < 1$.
4. El modelo de regresión no es autorregresivo, es decir no es del tipo $Y_j = \beta_1 + \beta_2 Y_{j-1} + \varepsilon_j$.
5. No faltan observaciones en los datos.

Para rechazar o no la hipótesis nula de que no hay autocorrelación de primer orden en las perturbaciones ε_j se consideran los límites inferior d_L y superior d_U, encontrados por Durbin y Watson, tales que si el valor *d* cae fuera de dichos límites existe posible presencia de correlación. La siguiente figura ilustra los criterios para el rechazo.



Rechazar H_0, evidencia de autocorrelación positiva

Rechazar H_0, evidencia de autocorrelación negativa

Si existe evidencia de autocorrelación es necesario buscar medidas remediales ya que aunque los estimadores de los coeficientes de regresión siguen siendo lineales, insesgados y consistentes bajo la presencia de autocorrelación, éstos no son eficientes (es decir, no tienen varianza mínima). Por lo tanto si se utiliza la $\text{var}(\hat{\beta}_2)$ la exactitud del estimador está inflada y al calcular $t_c = \dfrac{\hat{\beta}_2}{\sqrt{\text{var}(\hat{B}_2)}}$ se estará sobreestimando la significación estadística de $\hat{\beta}_2$.

Para hacer las correcciones primero es necesario estimar ρ (y esto se puede hacer mediante el procedimiento iterativo de Cochrane-Orcutt que se explicará más adelante) para hacer las siguientes transformaciones:

$$Y_j^* = Y_j - \hat{\rho} Y_{j-1} \tag{A.2}$$

$$X_j^* = X_j - \hat{\rho} X_{j-1} \tag{A.3}$$

y correr la regresión $Y_j^* = \beta_1^* + \beta_2^* X_j^*$ ($Y_j = \beta_1 + \beta_2 X_j$ es el modelo original). De esta forma se obtienen β_1^* y β_2^*, se prueba la significación de β_2^* y se comprueba qué tan bien se ajustan los datos transformados al nuevo modelo. Una vez hecho esto, es posible definir el nuevo modelo el cual tendrá la forma:

$$\hat{Y}_j = \frac{\beta_1^*}{1-\hat{\rho}} + \beta_2^* X_j . \tag{A.4}$$

Este procedimiento es iterativo, por lo que si el nuevo modelo sigue presentando autocorrelación, se hace una segunda regresión, una tercera y así sucesivamente hasta que los estimadores sucesivos de ρ difieran en menos de 0.01.

Anexo B Procedimiento iterativo de Cochrane-Orcutt para estimar ρ

Este procedimiento consiste en obtener $\hat{\rho}$ a partir del estadístico *d* de Durbin-Watson. Se basa en la estimación de los residuos e_j para obtener información sobre el valor desconocido de ρ.

Considérese que los términos de error estocástico ε_j se generan a través de un esquema autorregresivo de primer orden, i.e. $\varepsilon_j = \rho\varepsilon_{j-1} + u_j$. Los pasos a seguir son:

1. Estimar el modelo original y obtener los residuos e_j que son estimaciones de ε_j.
2. Con los residuos estimados correr la siguiente regresión: $e_j = \rho e_{j-1} + v_j$ para obtener $\hat{\rho}$ y poder hacer la corrección al modelo original.

Anexo C Ejemplo de corrección de autocorrelación: el caso de la zona MEDIA RESIDENCIAL de la división Occidente.

Se corrió el modelo Recíproco de la Edad para la zona MEDIA RESIDENCIAL. El tamaño de la muestra fue de 19 viviendas y los resultados desplegados por el paquete estadístico SPSS fueron los siguientes:

Model Summary(b)					
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**	**Durbin-Watson**
1	.739(a)	.545	.519	7.8106848916992E-02	**1.044**
a Predictors: (Constant), Recíproco de la edad					
b Dependent Variable: MEDIA RESIDENCIAL					

Según las tablas del estadístico *d* de Durbin-Watson a un nivel de significación de 0.05 se tiene que para una muestra de tamaño 19 $d_L = 1.180$ y $d_U = 1.401$. Como $d = 1.044$ en este ejemplo, existe evidencia de autocorrelación positiva.

ANOVA(b)						
Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
1	**Regression**	.124	1	.124	20.399	.000(a)
	Residual	.104	17	6.101E-03		
	Total	.228	18			
a Predictors: (Constant), Recíproco de la edad						
b Dependent Variable: MEDIA RESIDENCIAL						

Coefficients(a)						
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	.782	.044		17.748	.000
	Recíproco de la edad	-2.313	.512	-.739	-4.516	.000
a Dependent Variable: MEDIA RESIDENCIAL						

Se aprecia que los coeficientes son significativos a un nivel de significación de 0.05 pero es necesario corregir la autocorrelación pues se desean estimadores con varianza mínima. Por ello, se realizó el procedimiento iterativo de Cochrane-Orcutt y se estimó ρ utilizando los residuos del modelo original para correr la siguiente regresión:

$$e_j = \rho e_{j-1} + v_j \,.$$

(C.1)

Coefficients(a,b)						
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	residual anterior MEDIA RESIDENCIAL	**.431**	.329	.303	1.311	.207
a Dependent Variable: residual MEDIA RESIDENCIAL						
b Linear Regression through the Origin						

Con base en los resultados desplegados por SPSS se obtuvo que $\hat{\rho} = 0.431$. Una vez conocida esta estimación fue posible calcular las transformaciones:

$$D_j^* = D_j - \hat{\rho} D_{j-1} \tag{C.2}$$

$$\frac{1}{X_j^*} = \frac{1}{X_j} - \hat{\rho}\frac{1}{X_{j-1}} \tag{C.3}$$

y correr la regresión $D_j^* = \beta_1^* + \beta_2^* \frac{1}{X_j^*}$. Los resultados fueron los siguientes:

Model Summary(b)					
Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**	**Durbin-Watson**
1	.427(a)	.182	.131	7.531140592192E-02	1.381
a Predictors: (Constant), Recíproco de la edad*					
b Dependent Variable: MEDIA RESIDENCIAL*					

ANOVA(b)						
Model		**Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**
1	**Regression**	2.024E-02	1	2.024E-02	3.569	.077(a)
	Residual	9.075E-02	16	5.672E-03		
	Total	.111	17			
a Predictors: (Constant), Recíproco de la edad*						
b Dependent Variable: MEDIA RESIDENCIAL*						

Coefficients(a)						
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	.440	.053		8.358	.000
	Recíproco de la edad*	**-2.406**	1.274	-.427	-1.889	**.077**
a Dependent Variable: MEDIA RESIDENCIAL*						

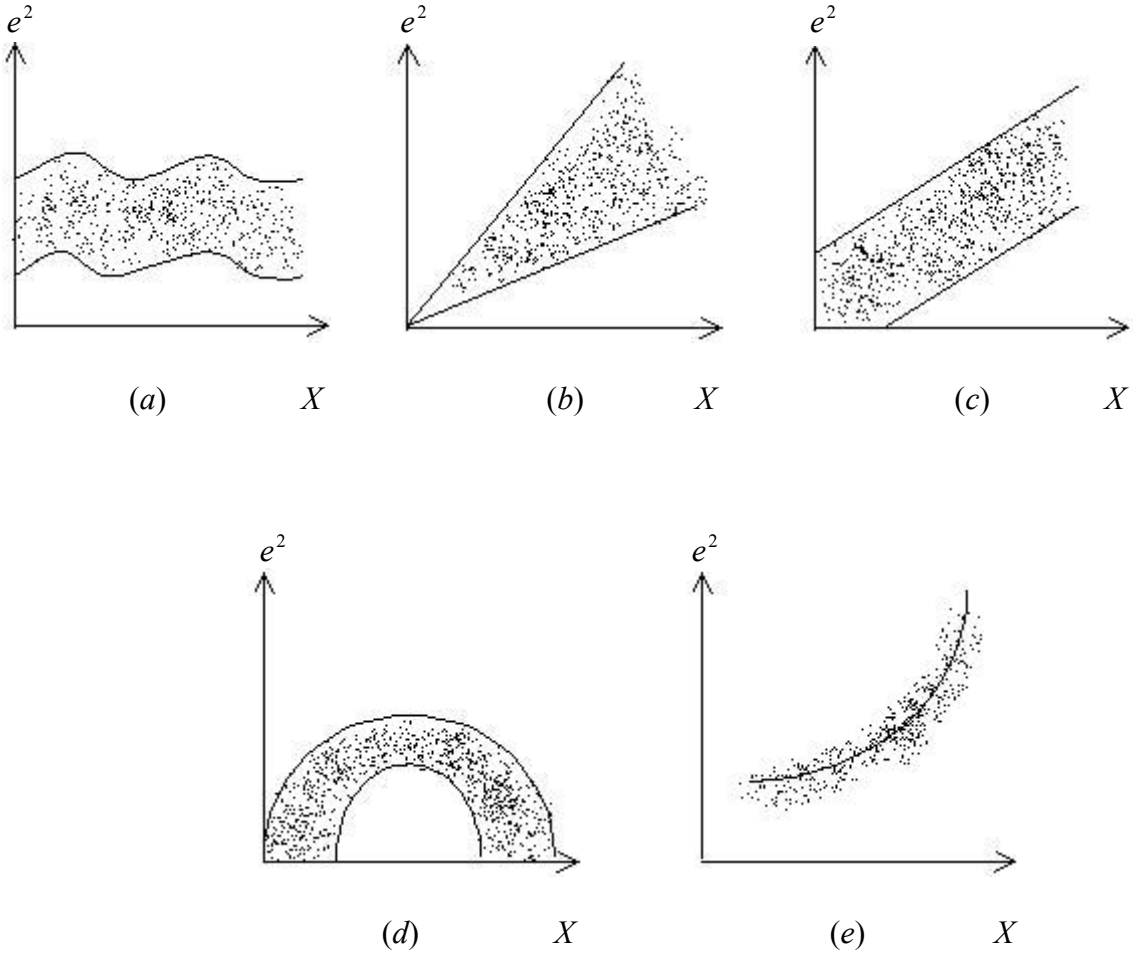
El número de observaciones se redujo a 18 por lo tanto $d_L = 1.158$ y $d_U = 1.391$. Como la *d* de Durbin-Watson resultó ser 1.381, este valor cae en la zona de indecisión, pero está muy cercano a la zona en donde se rechaza la presencia de autocorrelación. Por otro lado, la estimación de β_2^* no cumple exactamente con la condición de significación establecida como 0.05, pero la diferencia no es muy grande y en términos reales lo que estaría pasando es que se tiene una probabilidad mayor de considerar β_2^* como significativa cuando en realidad no lo sea.

Finalmente con estos resultados se obtuvo un nuevo modelo para la zona MEDIA RESIDENCIAL de la forma: $\hat{D}_j = \frac{\beta_1^*}{1-\hat{\rho}} + \beta_2^* \frac{1}{X_j}$; donde $\hat{\rho} = 0.431$, $\beta_1^* = 0.440$, $\beta_2^* = -2.406$, i.e.:

$$\hat{D}_j = \frac{-2.406}{X_j} + 0.773 . \qquad \text{(C.3)}$$

Anexo D Detección de la heterovarianza

Una manera simple para detectar la heterocedasticidad o heterovarianza cuando no existe información "*a priori"* es obteniendo un diagrama de dispersión de e_j^2 contra $\hat{Y}_j$ para ver si presentan algún patrón, pero para el caso del modelo de dos variables se pueden graficar los residuales al cuadrado contra la variable independiente *X*. A continuación se muestran algunos de los patrones que pudieran detectarse.



En el diagrama (*a*) se advierte que no existe un patrón sistemático entre las variables lo que sugiere la inexistencia de heterocedasticidad; en cambio, en los otros diagramas sí hay patrones definidos. Por ejemplo, el diagrama (*c*) sugiere una relación lineal y el (*e*) una relación cuadrática entre las variables.

Anexo E Prueba de Park

Park propone que σ_j^2 es una función de la variable independiente X_j. La forma funcional propuesta es:

$$\sigma_j^2 = \sigma^2 X_j^{\beta} e^{v_j}. \tag{E.1}$$

Si de (E.1) se obtiene el logaritmo natural se tiene:

$$\ln(\sigma_j^2) = \ln(\sigma^2) + \beta \ln(X_j) + v_j. \tag{E.2}$$

Dado que por lo general σ_j^2 es desconocida, se propone usar e_j^2 como una aproximación y se realice la siguiente regresión:

$$\ln(e_j^2) = \alpha + \beta \ln(X_j) + v_j. \tag{E.3}$$

Se prueba la hipótesis nula $H_0 : \beta = 0$. Si se rechaza H_0, $\beta \neq 0$ y entonces puede ser que exista heterovarianza.

Una vez que se detectó la presencia de heterovarianza es necesario corregirla ya que de no detectarla y corregirla, la $\text{var}(\hat{\beta}_2) = \dfrac{\sigma^2}{\sum_{j=1}^{m} x_j^2}$ que es la varianza común que se obtiene bajo el supuesto de varianzas iguales, es un estimador sesgado de la verdadera $\text{var}(\hat{\beta}_2)$. Esto implica que dependiendo de la naturaleza de la relación entre la varianza y los valores

que toma la variable independiente, en promedio se estará sobreestimando o subestimando la verdadera varianza de $\hat{\beta}_2$.

Para remediarla, se construye un nuevo modelo con base en la siguiente transformación:

$$\frac{Y_j}{X_j^{\beta/2}} = \frac{\beta_1}{X_j^{\beta/2}} + \beta_2 X_j^{(1-\beta/2)} + \frac{\varepsilon_j}{X_j^{\beta/2}} \tag{E.4}$$

donde β es un número que simboliza la relación entre los residuales al cuadrado y *X*. Por ejemplo, si la relación que sugiere el diagrama de dispersión es cuadrática una buena suposición del valor de β es 2. No obstante, pudiera ser que $\beta = 2$ no corrigiera el modelo entonces para ese caso deben probarse otros valores cercanos a dos. Si la relación sugerida es lineal, entonces $\beta = 1$ pudiera corregir la heterovarianza. Una vez que se construyó el nuevo modelo, se corre una regresión sobre éste y se vuelve a realizar al Prueba de Park.