

MODELOS PARA DATOS DE RECuento DE CORTE TRANSVERSAL CON EXCESO DE CEROS. APLICACIÓN A CITAS DE PATENTES

Margarita E. Romero Rodríguez. mromero@ull.es*

Departamento de Economía de las Instituciones, Estadística Económica y Econometría.
Universidad de la Laguna.

Enrique Los Arcos
Oficina Europea de Patentes. The Hague, Holanda

Víctor Cano Fernández. vcano@ull.es

Departamento de Economía de las Instituciones, Estadística Económica y Econometría.
Universidad de la Laguna.

Miguel Sánchez Padrón. msanchez@ull.es

Departamento de Economía Aplicada. Universidad de La Laguna.

Palabras clave: datos de recuento, Poisson, heterogeneidad no observada, sobredispersión, exceso de ceros, citas a patentes.

Resumen

En este trabajo se presentan y discuten diversos modelos de regresión para el análisis de datos de recuento (*count data*) de corte transversal, haciendo especial hincapié en la forma en que los distintos modelos recogen dos de las características más comunes de este tipo de datos, el exceso de ceros y la heterogeneidad no observada, cuyos efectos en el proceso de inferencia estadística son importantes.

Con la finalidad de poner de manifiesto las características de cada uno de estos modelos, se lleva a cabo su aplicación a un conjunto de datos sobre patentes, concedidas por la Oficina Europea de Patentes (OEP), que casi llegaron al final de su período legal de vigencia. En concreto se analiza una de las relaciones propuestas en la literatura, la planteada entre el número de citas recibidas por las patentes y el alcance de las mismas (*scope*), considerando, además, otro conjunto de características que pueden ser relevantes en la explicación del número de citas. La comparación de los resultados obtenidos con las distintas aproximaciones ilustra la influencia del tipo de modelo elegido sobre las conclusiones que pueden extraerse.

1. Introducción

Los datos de recuento se obtienen cuando se contabiliza el número de veces que ocurre un suceso en un intervalo de amplitud determinada y, en consecuencia, pueden ser considerados como realizaciones de una variable aleatoria que sólo toma valores enteros no negativos.

Los modelos de regresión estándar, como el modelo de regresión lineal, presentan claras deficiencias a la hora de ser utilizados para analizar este tipo de datos, ya que ignoran la especial naturaleza de la variable dependiente¹. Como consecuencia de ello, ha sido necesario definir nuevos modelos de regresión. El modelo de regresión de recuento por excelencia es el modelo de Poisson que, aunque ampliamente utilizado (Hausmann y otros (1984), Cameron y Trivedi (1986), entre otros), se apoya en fuertes supuestos distribucionales cuya rigidez puede dificultar la consecución de un buen ajuste. La búsqueda de mayor flexibilidad ha propiciado la aparición de otros modelos, algunos basados en la distribución de Poisson, que han recogido mejor algunas características a menudo presentes en los datos, tales como la sobredispersión, el exceso de ceros o la existencia de grandes colas a la derecha, consideradas como implicaciones de la heterogeneidad no observada (Mullahy, 1997).

Los primeros desarrollos significativos de los modelos para este tipo de datos tienen lugar en el ámbito de la ciencia actuarial, bioestadística y demografía, siendo posterior su aplicación al campo de la economía, ciencias políticas y sociología. Cameron y Trivedi (1998) afirman que la aparición de los modelos lineales generalizados, descritos inicialmente en Nelder y Wedderburn (1972) y entre los que destaca el modelo de regresión de Poisson, constituyó una pieza clave en el desarrollo de los modelos de regresión sobre datos de recuento.

Patil (1970) recoge numerosas aplicaciones del análisis de datos de recuento; sin embargo, el enfoque metodológico adoptado en la mayoría de éstas queda fuera del contexto de la regresión. También son muchas las aplicaciones de los modelos de

* Los autores agradecen los comentarios y sugerencias realizadas por el Dr. José Juan Cáceres Hernández.

¹ A no ser que la media de la variable dependiente sea elevada, en cuyo caso se puede utilizar la aproximación normal y métodos de regresión relacionados (Cameron y Trivedi, 1998).

regresión en el campo de la estadística y la econometría para este tipo de datos, no sólo de corte transversal sino también de series temporales o datos de panel².

En el presente trabajo se estudian diferentes modelos de regresión para el análisis de datos de recuento de corte transversal en los que se observa un exceso de ceros y se aplica cada uno de estos modelos al análisis del número de citas de patentes. Específicamente, se analiza la relación entre el número de citas recibidas por las patentes y el alcance o amplitud de la protección de las mismas.

En el siguiente epígrafe se analiza el modelo de regresión de Poisson, y se efectúa un breve repaso de los modelos de Poisson compuestos, con especial referencia al modelo binomial negativo. También se mencionan algunos modelos específicamente ideados para el análisis del exceso de ceros. En el epígrafe tercero se comentan algunas cuestiones sobre las citas de patentes y en el siguiente se describen los datos, se aborda la estimación de los modelos y se comentan los resultados obtenidos. Finalmente, se exponen las conclusiones del trabajo.

2. Modelos para datos de recuento

2.1. Modelo de regresión de Poisson

El modelo de referencia para el tipo de datos objeto de estudio es el *modelo de regresión de Poisson*. Se trata de un modelo no lineal, en el que el parámetro de intensidad (λ) del proceso de Poisson depende de un conjunto de variables explicativas. Si la relación entre el parámetro y los regresores es determinista, el modelo recibe el nombre de *modelo de regresión de Poisson estándar*, en cambio, cuando la relación es de naturaleza estocástica, se obtienen los denominados *modelos de Poisson mixtos*.

En estos últimos modelos se pretende explicar el número de veces que ocurre un suceso de interés para el individuo i ($i = 1, \dots, N$) en un intervalo de amplitud unitaria, como función de un vector de variables explicativas, X_i . En concreto, se supone que la

² Entre las aportaciones más interesantes cabe destacar los trabajos de Cameron y Trivedi (1986), Cameron y otros (1988), para modelizar el número de visitas médicas; Hausmann y otros (1984), para analizar patentes; Ozuna y Gómez (1995), para la demanda recreativa; Lambert (1992), para los defectos en manufacturas, etc.

distribución de $Y_i / X_i = x_i$ es Poisson de parámetro λ_i . La especificación estándar de este modelo viene dada por:

$$P(Y_i = y_i / x_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \quad y_i = 0, 1, 2, \dots$$

$$\lambda_i = E[y_i / x_i] = \exp(x_i' \beta)$$

Nótese que la formulación log-lineal³ del parámetro λ_i garantiza la naturaleza positiva de la media de la variable dependiente, donde β es un vector de parámetros y x_i una matriz de variables explicativas.

A pesar de su enorme popularidad, el modelo de Poisson presenta una serie de limitaciones importantes. Una de ellas y al mismo tiempo una de sus principales características, es la igualdad de la media y varianza condicionales, conocida por *equidispersión*. Dicha propiedad implica la naturaleza heterocedástica del modelo e impide captar la sobredispersión que está presente en la mayoría de los datos habitualmente utilizados y, por tanto, dificulta la obtención de un buen ajuste. En este modelo la única fuente de diferencias entre los individuos es atribuida a los distintos valores de las variables explicativas; sin embargo, puede haber diferencias de otro tipo, que normalmente tratan de ser recogidas introduciendo un término de heterogeneidad en el modelo. Mullahy (1997) argumenta que esta heterogeneidad no observada conlleva sobredispersión y exceso de ceros. Otra de las limitaciones es el supuesto de independencia propio de un proceso de Poisson, en el que se considera que la ocurrencia de un suceso en un determinado intervalo no modifica la probabilidad de ocurrencia de otro suceso en otro intervalo no solapado. El incumplimiento del supuesto de independencia puede ser la causa de la sobredispersión (Winkelmann, 1995). Finalmente, y en relación con lo comentado anteriormente, el modelo de Poisson se presenta como un modelo demasiado sencillo para captar el exceso de ceros que está presente en muchos datos, esto es, subestima la frecuencia real de ceros, pero además, tiende a sobrestimar la frecuencia real de valores pequeños y a subestimar la de valores elevados.

³ A esta expresión se le suele denominar *función media exponencial*.

2.2. Modelos de Poisson compuestos

El problema de la heterogeneidad no medida surge en aplicaciones en las que las diferencias de comportamiento entre individuos no pueden ser adecuadamente capturadas por el conjunto de variables explicativas de la función media condicional del modelo.

En este sentido, Gurmu y otros (1999) señalan que la heterogeneidad no observada es sumamente importante en modelos económicos que utilizan datos de recuento. De hecho, muchos modelos de regresión no lineales son muy sensibles a la omisión de la misma. Una de sus consecuencias más importantes es la sobredispersión⁴, afectando, asimismo, a la estructura de los momentos de órdenes más bajos de la variable dependiente del modelo. Además, y como destaca Mullahy (1997) tiene, otras dos implicaciones, el exceso de ceros y la existencia de colas amplias.

Habitualmente, esta heterogeneidad no observada se recoge introduciendo un término de error multiplicativo en la media condicional del modelo de Poisson, dando lugar, así, a los *modelos de Poisson mixtos o compuestos*, donde

$$\lambda_i^* = E[y_i / x_i, v_i] = \lambda_i \quad v_i = e^{x_i' \beta} e^{\varepsilon_i}$$

El término de heterogeneidad no observada, v_i , puede recoger un error de especificación, como la omisión de alguna variable explicativa (Gourieroux y otros, 1984a, b) o bien la aleatoriedad intrínseca del proceso (Hausmann y otros, 1984). Normalmente, se supone que las v_i se distribuyen idéntica e independientemente con una distribución paramétrica conocida y son independientes del conjunto de variables explicativas. También suele suponerse que su media es la unidad y su varianza $\sigma_{v_i}^2$.

Si definimos $g(v_i)$ como la función de densidad de probabilidad para v_i , la distribución de probabilidad marginal de Y_i puede ser obtenida integrando respecto a v_i . Si se supone, además, que la variable $(Y/x, v)$ sigue una distribución de Poisson de parámetro λ_i^* , la distribución marginal de Y_i queda

⁴ Hausman y otros (1984), Cameron y Trivedi (1986).

$$P(Y_i = y_i / x_i) = \int P(Y_i = y_i / x_i, v_i) g(v_i) dv_i = \int \frac{e^{-\lambda_i^*} \lambda_i^{*y_i}}{y_i!} g(v_i) dv_i$$

Esta expresión define una *distribución de Poisson compuesta* cuya forma precisa depende de la elección específica de $g(v_i)$. Por otro lado, no es necesario especificar $g(v_i)$ para mostrar la existencia de sobredispersión

$$V[y_i / x_i] = \lambda_i (1 + \sigma_{v_i}^2 \lambda_i) > E[y_i / x_i] = \lambda_i$$

En la literatura se han usado varias parametrizaciones de $\sigma_{v_i}^2$, algunas de las cuales han dado lugar a los llamados *modelos Negbin* (binomial negativo).

2.2.1. Modelo binomial negativo

Aunque el modelo binomial negativo puede ser motivado de diferentes formas⁵, la representación más común es considerar a éste como un modelo de Poisson compuesto. Tal representación se consigue bajo el supuesto de que el término de heterogeneidad no observada, v_i , se distribuye como gamma ($\Gamma(\delta, \delta)$) con $\sigma_{v_i}^2 = 1/\delta \equiv \alpha$, parámetro de dispersión, lo que conduce a la distribución de probabilidad binomial negativa

$$P(Y_i = y_i / x_i) = \frac{\Gamma(\alpha^{-1} + y_i)}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i} \right)^{\alpha^{-1}} \left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i} \right)^{y_i}$$

con media y varianza de la forma,

$$E[y_i / x_i] = \lambda_i$$

$$V[y_i / x_i] = \lambda_i (1 + \alpha \lambda_i)$$

Cameron y Trivedi (1986) destacan que la consideración de

$$\alpha^{-1} = \frac{1}{\theta} \left(e^{x_i \beta} \right)$$

con θ positivo y k una constante arbitraria, permite obtener un amplio rango de relaciones media-varianza

⁵ Ver Boswell y Patil (1970).

$$V[y_i / x_i] = \lambda_i (1 + \alpha \lambda_i) = e^{x_i' \beta} + \theta \left(e^{x_i' \beta} \right)^{2-k} = E[y_i / x_i] + \theta E[y_i / x_i]^{2-k}$$

que, a su vez, permite hablar de distintos modelos NEGBIN (NEGBIN I si $k=1$ y NEGBIN II si $k=0$, entre otros).

2.3. Modelos para el exceso de ceros

La abundancia de ceros en los datos ha sido analizada a partir de diferentes aproximaciones. Las más destacadas son las propuestas por Mullahy (1986), que introdujo el llamado modelo de Poisson con obstáculo, y Lambert (1992), que ideó el modelo de Poisson inflado con ceros (ZIP).

Debido a que la aproximación Poisson con obstáculo no ha sido utilizada en la aplicación que se realiza, sólo haremos referencia al modelo de Lambert. En este modelo se postula que existe una probabilidad p de observar el resultado cero y una probabilidad $(1-p)$ de que el valor observado proceda de una distribución de Poisson.

Es decir, su distribución de probabilidad viene dada por

$$P(Y_i = y_i) = \begin{cases} p_i + (1 - p_i) e^{-\lambda_i} & \text{si } y_i = 0 \\ (1 - p_i) e^{-\lambda_i} \lambda_i^k / k! & \text{si } y_i = k > 0 \end{cases}$$

donde los vectores de parámetros λ_i y p_i satisfacen las siguientes condiciones

$$\begin{aligned} \log(\lambda_i) &= Z_i \beta \\ \log it(p_i) &= \log \left(\frac{p_i}{1 - p_i} \right) = W_i \gamma \end{aligned}$$

siendo Z_i y W_i vectores de variables explicativas, que podrían o no coincidir.

Las diferentes formas en que los parámetros λ y p pueden ser especificados dan lugar a diferentes versiones del modelo⁶.

⁶ Otras aproximaciones para modelizar el exceso de ceros pueden verse en Cohen (1963), Heilbron (1989) y Farewell (1986), entre otros.

3. Citas, reivindicaciones y amplitud de las patentes⁷

Las citas que aparecen en los documentos de patentes han sido utilizadas con diversos propósitos. De forma muy general, los trabajos que hacen uso de éstas, se han centrado en dos cuestiones: a) estudiar la naturaleza y dirección de las externalidades del conocimiento, y b) inferir la calidad o importancia de la invención citada, (Hall y otros, 1999:6).

En este segundo grupo cabría ubicar el trabajo de Lerner (1994) en el que se examina la relación entre el valor de las empresas y la amplitud de las patentes concedidas a éstas. Este autor, utiliza la relación encontrada entre el número de citas y la amplitud de las patentes para validar la variable que utiliza como medida de ésta.

Algunos trabajos recientes han intentado validar las citas de patentes como medida del valor económico de las empresas. Harhoff y otros (1999), encuentran una relación positiva entre estas variables pero bastante *ruidosa*. Asimismo, Hall y otros (1999), haciendo referencia a algunos trabajos, destacan que “todos esos autores han encontrado que las patentes están correlacionadas con el valor económico, y algunos encuentran que el número de citas está aún más correlacionado” (Hall y otros, 1999:2), aunque reconocen el sustancial “ruido” en los datos de las citas. De hecho, “las investigaciones conducidas por Schmoch (1993) y Narin y otros (1988) revelan que una patente que es ampliamente citada no es siempre una patente importante; y, aunque tiende a ser verdad que una patente importante sistemáticamente recibe un gran número de citas, esto no es necesariamente siempre así” (Joly y de Looze, 1996:1027).

En este trabajo se utiliza un conjunto de patentes que han sido mantenidas hasta los 19 años (casi el máximo periodo legal de 20 años), con el objetivo de estudiar la relación entre el número de veces que una patente es citada, como índice de su importancia, y la amplitud (*scope*) de las mismas. Sin embargo, antes de abordar la estimación de esta relación, es necesario realizar un breve comentario sobre la naturaleza y papel de las reivindicaciones (*claims*) en las patentes.

⁷ Una discusión detallada de éstos y otros aspectos relacionados con los sistemas de patentes, puede verse en Sánchez-Padrón y otros (1999b).

Una reivindicación en una patente describe su área de protección y está compuesta de un cierto número de características técnicas⁸. Puesto que las reivindicaciones son el principal determinante de la amplitud del monopolio de una patente, la interpretación de éstas es frecuentemente el elemento esencial de un litigio. Tal interpretación sobre la amplitud de la protección puede situarse entre los dos extremos siguientes: restringir la amplitud de la protección a la literalmente recogida en la(s) reivindicación(es)- esta es la aproximación rigurosa de Gran Bretaña - o bien, ampliar la protección para incluir todo lo que usa el centro de la idea inventiva expresada en la reivindicación de la patente. “A la luz de la influencia de la regla europea sobre la amplitud de la protección, y también teniendo en cuenta el hecho de que muchas legislaciones de patentes no contienen disposiciones sobre este aspecto, no es difícil entender el efecto de la regulación en Europa. Además, es concisa y claramente redactada...” (Stauder, 1992:23). De hecho, la amplitud de la protección citada en el artículo 69 (1) del EPC (European Patent Convention) “fue desde el comienzo [del trabajo preparatorio del EPC] una solución de compromiso que es recogida en la última sentencia del Protocolo sobre Interpretación: ‘[La amplitud de la protección] debe ser interpretada como la definición de una posición entre esos extremos que combina una protección justa para el que patenta con un grado razonable de protección para las terceras partes’”. (Stauder, 1992:327).

En este sentido, las citas en las patentes cumplen una importante función legal, determinan el alcance (*scope*) y contenido de la técnica anterior (*prior art*), que es utilizado para averiguar las diferencias entre la invención reclamada y las anteriores. Las citas sirven, por tanto, para valorar la originalidad (novedad) de las reivindicaciones y para delimitar el contorno de esa originalidad (inventiva)⁹. Siguiendo a algunos de los principales autores en el análisis de citas de patentes, la idea básica detrás de esta aproximación es muy simple: “si una patente U.S. es citada por un examinador en muchas patentes U.S. siguientes, entonces la primera patente altamente citada ha sido la

⁸ El propósito de las reivindicaciones en la patente es delimitar el alcance del monopolio. El artículo 85 de la EPC establece el criterio que éstas deben cumplir: “Las reivindicaciones *definirán* el objeto para el cual se solicita la protección. Ellas serán *claras y concisas* y estarán apoyadas por la *descripción*”. En conexión con ese criterio, es importante destacar que de acuerdo al artículo 82, “La solicitud de una patente europea (y por consiguiente las reivindicaciones) relatarán una única invención o un grupo de invenciones relacionadas para formar un concepto general inventivo”. Esto es, “dos invenciones no se tendrán por el precio de una” (Cornish, 1997:129).

técnica anterior para muchas patentes posteriores, y es probable que contenga un avance significativo que ha precedido a esas subsecuentes invenciones” (Albert y otros, 1991:251). Sin embargo, es importante destacar que algunas cuestiones complejas acechan detrás de este velo de aparente simplicidad. En primer lugar, se pone en duda la validez del análisis de citas en general¹⁰. En segundo lugar, está la cuestión de las prácticas de los examinadores en relación a las citas¹¹. En tercer lugar, y quizás más importante, la validez de un índice basado en citas de patentes requiere un juicio externo para valorar la importancia de la invención patentada¹².

4. Datos y Resultados

4.1. Datos

Como ya se mencionó en el apartado anterior, en este trabajo se utilizan los datos referidos a un conjunto de patentes que han sobrevivido 19 años, casi el final de su vida legal. Estos datos han sido obtenidos a partir de la información suministrada por las Oficinas de Patentes de Francia y Holanda, sobre patentes europeas para las que habían sido pagadas las tasas de renovación hasta el decimonoveno año¹³. El número total de patentes es de 346 para Francia, 447 para Holanda y 164 comunes a ambos

⁹ Una diferencia muy importante entre el procedimiento de la OEP y el de otras oficinas de patentes es que todos los documentos citados en el informe de búsqueda están identificados con una letra en la primera columna de la hoja de citas (Véase Sánchez-Padrón y otros, 1999b: Anexo)

¹⁰ Para nuestra propuesta, necesitamos mencionar “la recurrencia frecuente de la misma clase de críticas sobre el uso de las citas”, y que “a pesar de las frecuentes y a menudo extremadamente perjudiciales críticas, el uso de las citas continúa” (Woolgar, 1991:320).

¹¹ Por ejemplo, Collins y Wyatt hacen la siguiente sorprendente afirmación sin ningún apoyo en la evidencia: “los examinadores tienden a restringir su lectura a un rango estrecho de especialidades y no están familiarizados con una literatura más amplia. Cuando ellos usan la literatura de las revistas tiende a ser en una forma secundaria (revistas de abstracts) más que de forma primaria, y las repeticiones ocasionales de errores bibliográficos sugieren que los examinadores utilizan citas de artículos que no han sido leídos de primera mano. La utilización de los examinadores del mismo conjunto de citas en muchos patentes diferentes sugieren una tendencia ocasional a citar por rutina más que por relevancia” (Collins y Wyatt, 1988:67). También, algunos comentarios al respecto pueden encontrarse en Meyer (2000).

¹² Esta es probablemente la razón por la que los estudios de citas han intentado diferentes formas para validar la aproximación, esto es, contrastar que las patentes más citadas son patentes de gran impacto o importancia; por ejemplo, preguntando a expertos sobre su valoración mediante ciertos indicadores tecnológicos, véase, por ejemplo, (Albert y otros, 1991).

¹³ Una vez concedidas, las patentes europeas pasan a los estados designados y pagan las tasas de renovación en los sistemas nacionales. Las estadísticas de esas renovaciones no están disponibles en la base de datos on-line de la OEP. Puesto que, dependiendo de las leyes y reglas nacionales – son posibles retrasos o adelantos de los pagos anuales- las estadísticas de renovación pueden mostrar divergencias. Por eso, nosotros usamos la expresión “*habiendo pagado las tasas de renovación*”, aunque en el texto nos referiremos a esas patentes como de 19 años.

países, siendo el total de 629. El proceso de recopilación de los datos nos llevó a obtener la siguiente información para cada una de las patentes¹⁴:

- 1) APD: Fecha de Solicitud
- 2) GRD: Fecha de Concesión
- 3) DST: Estados Designados
- 4) IPC: Clasificación Internacional de Patentes
- 5) PRD: Fecha de Prioridad
- 6) APN: Nombre del Solicitante
- 7) TI: Título
- 8) CT: Citas

A partir de esta información se construyeron las variables cuyo resumen descriptivo se muestra en la tabla 1 del anexo. Un primer aspecto que conviene destacar es que, aunque el año de solicitud de esas patentes, 1978, se corresponde con el inicio de la OEP, las estadísticas referidas a esas patentes, en general, no muestran rasgos muy divergentes respecto cifras más actuales de la OEP. Donde se encuentran algunas diferencias es, en primer lugar, en relación al periodo transcurrido hasta la concesión de las mismas. Así, para nuestro grupo de patentes, la fecha media de concesión es aproximadamente, 1982, y el porcentaje medio de patentes pendientes de concesión después de tres años es del 52,6% (no mostrado en la tabla), que es similar al porcentaje de concesiones pendientes, solicitadas en 1980 (49.80%). En cambio, para las patentes solicitadas en 1987, éste se incrementa hasta el 77.7%¹⁵. Esto puede ser explicado por la ausencia de solicitudes pendientes en el año de inicio de la OEP.

En segundo lugar, el número medio de estados designados (DS) por patente es de 6,3, mientras que la media para 1996 es de 7,11. Esta diferencia se corresponde con la tendencia creciente que ha marcado la evolución de esta variable desde 1987. Ese año el número medio de estados designados fue de 6,25¹⁶.

¹⁴ En este proceso, los datos fueron completados utilizando diversas bases de datos.

¹⁵ Véase Sánchez-Padrón y otros (1999a).

¹⁶ Una cuestión que merece un análisis aparte es que el número medio de estados designados por las patentes de empresas japonesas es mucho más bajo y casi constante, de 4,31 en 1987 a 4,51 en 1996.

Por otra parte, el número medio de clases a cuatro dígitos del IPC (I2) es de 1,7075 con una desviación estándar de 0.8296. Esta variable es la que utiliza Lerner (1994) como medida de la amplitud de la patente. Específicamente, para la construcción de esta variable se considera que una patente asignada a las clases C07C61/63, C07C69/743 y C07D209/48 es contada como de dos clases¹⁷. También hemos considerado el número de clases completas (I1), esto es, sin restringir el número de dígitos, en el caso anterior el valor de la variable sería de tres, como es evidente la media de esta variable es mayor (2,3831).

Como hemos indicado, algunos trabajos establecen una relación directa entre el número de citas recibidas y la importancia de las patentes. Una primera aproximación para valorar la importancia del conjunto de patentes incluidas en nuestra muestra es examinar el número de citas recibidas de las patentes solicitadas posteriormente¹⁸. Como se observa en la tabla 1, el número medio de citas recibidas (C) por las patentes es de 2,3 con una desviación estándar de 3.37; el ratio varianza-media para esta variable es 4.77, lo que indica la presencia de una marcada sobredispersión en los datos.

En la tabla 2 del anexo se muestra la distribución de frecuencias de las citas para las 629 patentes de la muestra. Es interesante destacar que la distribución de las citas muestra una patrón muy asimétrico, sólo el 4,6% de las patentes han recibido 10 o más citas, y un 84% han recibido 5 o menos¹⁹. La distribución de frecuencias de nuestras patentes de 19 años, presenta la típica distribución asimétrica encontrada en poblaciones de patentes de mayor tamaño.

Sin embargo, las frecuencias encontradas en nuestra muestra son mucho más bajas que las obtenidas para EE.UU.²⁰, donde “la frecuencia media de citas para todas las patentes US está sobre 5 citas por patente, y sólo el 3% de las patentes son citadas más de 20 veces en 20 años” (Breitzman y Narin, 1996:10). Por otra parte, en la literatura abundan referencias relativas al cuidado con el que deben ser utilizadas las

¹⁷ Aunque no se muestra aquí, conviene poner de manifiesto que el 46,3% de las nuestras patentes se concentran en 12 clases de este tipo.

¹⁸ Debe tenerse en cuenta que no todas las patentes citantes han sido concedidas.

¹⁹ De esta forma, suponiendo que el número de citas a una patente puede ser indicativo de la importancia del “*avance tecnológico*” contenido en esa patente, los modelos de renovación de patentes no permiten mostrar las patentes más citadas y por consiguiente las patentes que contienen importantes avances tecnológicos.

²⁰ Un comentario a este respecto puede verse en Harhoff y otros (1999) p. 514.

tasas de citas. A modo de ejemplo podemos mencionar las siguientes: “el mero hallazgo de una gran frecuencia de citas ni supone ni debe suponer nada en sí mismo” (Trajtemberg, 1990:174). “Nosotros también encontramos que ciertas patentes muy citadas fueron abandonadas después de ser mantenidas durante algún tiempo. En realidad, aunque las citas a una patente está correlacionadas con la calidad de la patente, uno tiene que ser extremadamente cuidadoso cuando utiliza este método para identificar ‘patentes importantes’” (Joly y de Looze, 1996:1037). Además, la referencia de lo que constituye una alta o baja tasa de citas, no sólo varía entre diferentes estudios, sino a veces presenta un rango flexible²¹.

4.2. Estimación de los modelos y resultados

Para el análisis del impacto de la amplitud (*scope*) sobre el número de citas, se utilizan las diferentes aproximaciones presentadas en el epígrafe segundo. En este análisis, la variable dependiente es el número de citas recibidas por cada una de las patentes de la muestra (C) y las variables explicativas son DS, I2, GR1²² y un conjunto de ocho variables cualitativas que recogen las distintas clases del IPC.

En el modelo de Poisson las estimaciones de los coeficientes son todas significativas, a excepción de las de las dummies D4, D5, D7 y D8. Por otro lado, sus signos son los esperados, exceptuando el del coeficiente de la variable GR1. Concretamente, existe una relación positiva entre el número de citas y el número de clasificaciones a cuatro dígitos (I2), pudiendo interpretarse el coeficiente de esta variable como el incremento proporcional en el número esperado de citas ante un incremento en una clase a 4 dígitos, que en este caso sería del 22%, inferior al obtenido en el NBII, que es del 25%. También es interesante destacar que el coeficiente del número de estados designados (*¿proxy del tamaño esperado del mercado?*) es positivo, pero más bajo que el de I2 (en todos los modelos). *¿Puede este resultado ser*

²¹ Por dar algunos ejemplos: “Sobre el nivel de 5 citas por patentes pueden identificarse 53 patentes importantes, esto es, el 5% del corpus completo [de patentes en biotecnología para plantas durante 1980]. Esto se corresponde con una definición restrictiva de patentes importantes (en otros estudios, las patentes importantes representan el 10% (Schmoch (1993)))” (Joly y de Looze, 1996:1034). “La idea clave detrás del análisis de citas a patentes es que cuando una patente es altamente citada, esto es, citadas en 5, 10, 20 ó más patentes subsecuentes, entonces esa patente es probable que contenga un importante avance tecnológico, un avance sobre el que muchas patentes posteriores se construyen” (Narin, 1993: 26).

²² Esta variable mide el número de años existente entre las fechas de solicitud y concesión de la patente y se ha introducido en el modelo para medir el efecto de las diferencias de tiempo para acumular citas a partir de patentes posteriores.

interpretado como indicativo de que las citas son explicadas mejor por variables relacionadas con la tecnología que por variables económicas? En conjunto, las variables que controlan las clases del IPC son significativas²³. Para la variable GR1, *a priori*, podría esperarse la existencia de una relación negativa con el número de citas, indicando que cuanto mayor es el tiempo que tarda en ser concedida una patente desde que es solicitada, menor debería ser la probabilidad de ser citada, sin embargo, para los datos disponibles el signo resultante es positivo, lo cual puede ser evidencia de que algunas patentes “importantes” han tenido un período más dificultoso desde la solicitud a la concesión²⁴.

Con el fin de captar la sobredispersión presente en los datos se ha incluido un término aleatorio en la media condicional de la variable dependiente, que sigue una distribución gamma y se han hecho dos supuestos sobre la función varianza. Por un lado se ha supuesto que la varianza es proporcional a la media, lo que ha dado lugar al modelo Negbin I (NBI), y por otro lado se ha definido una varianza que es función cuadrática de la media, resultando así el modelo Negbin II (NBII). En ambos modelos, aunque los signos de los coeficientes no cambian respecto al modelo de Poisson, si que se observan, en general, cambios notables en la significación de éstos: el coeficiente del número de estados designados ya no es significativo en el NBI y sigue siéndolo en el NBII, pero al 10%. En ambos modelos, las únicas dummies con coeficientes significativos son D2 y D6. Estos cambios en la significación de los coeficientes parecen coherentes con la idea de que al no estar recogida la sobredispersión de los datos en el modelo de Poisson, la varianza de las estimaciones resulta menos elevada, lo que tiende a inflar los valores de los estadísticos de significación individual y, por tanto, a sobrevalorar la influencia de los regresores. Además, el parámetro de dispersión, α , es claramente significativo en los dos modelos, como se muestra en la tabla 3 del anexo.

Para comparar el modelo de Poisson con las aproximaciones binomiales negativas se han utilizado varias herramientas estadísticas. Se han comparado las funciones de log-verosimilitud en los tres casos (-1614.6, -1268.9 y -1255.9,

²³ En la ecuación de regresión estimada, la categoría de referencia fue la clase A del IPC.

²⁴ Para una discusión detallada de las 29 patentes más citadas de nuestra muestra véase Sanchez Padrón y otros (1999b). Un dato significativo es que la media del período de concesión para estas patentes es

respectivamente) y se ha calculado el CAIC para cada uno²⁵ (3311, 2612, 2586) resultando favorecido en todos los casos el modelo NBII seguido del NBI. Por otro lado, se ha utilizado el *test óptimo basado en la regresión* propuesto por Cameron y Trivedi (1990) para contrastar la sobredispersión o subdispersión en el modelo de Poisson²⁶. Este test se basa en la regresión auxiliar MCO de $z_i = \left[(y_i - \mu_i)^2 - y_i \right] \sqrt{2} \mu_i$ sobre $w_i = g(\mu_i) / \sqrt{2} \mu_i$, donde $g(\mu_i)$ es igual a μ_i o μ_i^2 y en el posterior análisis de la significación del coeficiente de la misma. En ambos casos, dicho coeficiente resultó positivo y significativo, indicando la presencia de sobredispersión en los datos y el consiguiente rechazo del modelo de Poisson.

También, y como se muestran en la tabla 4 del anexo, se emplearon los tests de habituales de sobredispersión que conducen al claro rechazo del modelo de Poisson frente al modelo NBI y al NBII.

Finalmente, se utilizó un modelo ZIP, en el que la probabilidad de la observación cero se modelizó suponiendo una distribución logística para v_i , de la forma,

$$v_i = \tau \ln(\lambda_i)$$

siendo τ un nuevo parámetro a estimar.

En este modelo, todos los coeficientes son significativos con las excepciones de los de las variables D4 y D8, además, los signos se mantienen iguales que en los anteriores. El parámetro τ también es significativo y negativo.

Hemos comparado el modelo de Poisson con el ZIP utilizando el estadístico de Vuong (1986), propuesto para contrastar modelos no anidados. El valor del estadístico es de 8.11, indicando, nuevamente, el claro rechazo del modelo de Poisson.

mayor que la del conjunto y además un 20% de ellas fueron litigadas, que puede confirmar la importancia de las patentes de esta submuestra (Véase, Lerner (1994)).

²⁵ El CAIC, comparado con el AIC usual, impone una penalización relacionada con el tamaño de la muestra y es asintóticamente consistente.

²⁶
$$\begin{cases} H_0 : \text{var}(y_i) = \mu_i \\ H_1 : \text{var}(y_i) = \mu_i + \alpha g(\mu_i) \end{cases}$$

Por último, también hemos calculado las frecuencias ajustadas para cada modelo (tabla 5 del anexo). Al comparar las frecuencias observadas y las ajustadas para el modelo de Poisson se pone de manifiesto lo ya mencionado para este caso, que subestima el número de ceros, así como los recuentos por encima de 6. Respecto a los modelos Negbin se observa un doble comportamiento, por un lado, el pobre ajuste obtenido por el NBI, inflando, en exceso, el número de ceros lo que lleva a subestimar el resto de recuentos, y por otro lado, el aparente buen comportamiento del NBII no sólo para el recuento cero sino también para el resto de recuentos. El modelo ZIP, como era de esperar, recoge mejor el número de ceros que el modelo de Poisson, pero muestra distorsiones para los recuentos de 1 a 4. Estos aspectos se confirman observando los resultados del test χ^2 de bondad de ajuste y comparando las medias y varianzas estimadas de los distintos modelos con las observadas.

Adicionalmente, con el fin de evaluar la capacidad de los modelos cuando el número de ceros es aún mayor, y considerando el hecho, a veces señalado en la literatura sobre patentes²⁷, de que en muchos documentos de patentes las citas que aparecen hacen referencia meramente a antecedentes tecnológicos y no tienen una fuerte relación con la patente examinada, se decidió agrupar los recuentos nulos y los recuentos iguales a la unidad para el número de citas, de forma que el número de ceros de la muestra pasó de 196 (31%) a 345 (55%) y el resto de recuentos se redujo en una unidad. En líneas generales, las conclusiones respecto a cada modelo y también respecto a la comparación de modelos, se siguen manteniendo para los datos modificados. Sin embargo, es interesante resaltar que, analizando nuevamente las frecuencias ajustadas, el modelo ZIP recoge mejor este mayor número de ceros, así como el resto de recuentos, que con las observaciones originales.

En cualquier caso, debe puntualizarse que, si bien, algunos de los modelos captan apropiadamente el comportamiento del número de citas - revelando, al mismo tiempo, la relación encontrada por Lerner (1994) - las medidas de bondad muestran - como ya se ha apuntado en el apartado 3 y también en muchos trabajos que hacen uso de las citas de patentes - que la relación es de carácter *ruidoso*. Además, conviene no perder de vista que las patentes aquí consideradas llegaron casi al final de su periodo

²⁷ A este respecto puede consultarse, entre otros, Albert y otros (1991), Sánchez-Padrón (1999b) y Meyer (2000).

legal de vigencia, lo que podría ser indicativo, por sí mismo, de la importancia de éstas patentes²⁸. Sin embargo, las características recogidas ponen de manifiesto una gran heterogeneidad de éstas y, de hecho, podría afirmarse que cada patente es un mundo en sí misma: "los sistemas de patentes protegen una amplia variedad de novedades técnicas, desde grandes avances, que crearán nuevas industrias, hasta pequeñas mejoras en productos establecidos" (Cornish, 1997:110). En cualquier caso, parece que sólo unas pocas patentes son considerablemente significativas, mientras que el resto protegen innovaciones incrementales de menor entidad²⁹.

5. Conclusiones

En este trabajo se han descrito diferentes modelos de regresión para el análisis de datos de recuento de corte transversal, en los que se considera el exceso de ceros y la heterogeneidad no observada. Con objeto de ilustrar la utilidad de estos modelos se ha elegido como ámbito de aplicación el análisis de la relación entre el número de citas recibidas por las patentes y el alcance o amplitud de las mismas. Los resultados obtenidos permiten concluir, en primer lugar, que el modelo de Poisson proporciona un pobre ajuste puesto que se muestra incapaz de recoger la marcada sobredispersión presente en los datos y el considerable número de ceros. En cambio, el modelo Negbin II se adapta mejor a ambas características, mientras que el ZIP sobreestima el número de ceros. En segundo lugar, en todos los modelos se obtiene una relación positiva y significativa entre el número de citas y la variable utilizada como medida de la amplitud - el número de clases a cuatro dígitos del IPC - y también respecto al número de estados designados, aunque en éste último caso con un menor efecto. Si pudiéramos considerar esta variable como una proxy del tamaño esperado del mercado, cabría preguntarse si el resultado obtenido puede ser interpretado como indicativo de que las citas son explicadas en mayor medida por variables relacionadas con la tecnología que por variables económicas.

En cualquier caso, debe destacarse que la muestra de patentes utilizada en este trabajo se corresponde con las patentes solicitadas en un año que podríamos considerar

²⁸ Aunque la relación entre el pago de las tasas de renovación y la importancia económica de las invenciones patentadas no es necesariamente sencilla, dichos pagos han sido usados como ponderaciones para construir índices del valor de las ideas patentadas.

²⁹ La fuerza de este argumento es aún mayor si observamos que la gran mayoría de las patentes no son mantenidas hasta el final de su periodo legal de vigencia, aunque éste no es nuestro caso.

inusual, el inicio de la OEP. Por otro lado, las características descritas sobre estas patentes ponen de manifiesto su gran heterogeneidad, por lo tanto, es difícil elaborar generalizaciones significativas, y se justifica la necesidad de efectuar análisis más precisos.

6. Bibliografía

Albert, M.B., Avery, D., Narin, F. y McAllister, P. (1991), "Direct Validation of Citation Counts as Indicators of Industrially Important Patents", *Research Policy*, 20, 251-259.

Boswell, M.T. y Patil, G.P. (1970), "Chance Mechanisms Generating the Negative Binomial Distributions", in Patil, G.P., ed. *Random Counts in Models and Structures*, vol 1-3, University Park, PA, and London, Pennsylvania State University Press.

Breitzman, A. y Narin, F. (1996), "A Case for Patent Citation Analysis in Litigation", *The Law Works*, March.

Cameron, A.C. y Trivedi, P.K. (1986), "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators", *Journal of Applied Econometrics*, 1, 29-53.

Cameron, A.C. y Trivedi, P.K. (1990), "Regression-based Tests for overdispersion in the Poisson Model", *Journal of Econometrics*, 46, 347-364.

Cameron, A.C. y Trivedi, P.K. (1998), "Regression Analysis of Count Data", *Cambridge University Press*.

Cameron, A.C., Trivedi, P.K., Milne, F. y Piggott, J. (1988), "A Microeconomic Model for the Demand for Health Care and Health Insurance in Australia", *Review of Economics Studies*, 55, 85-106.

Cohen, A.C. (1963), "Estimation in Mixtures of Discrete Distributions", in *Proceedings of the International Symposium on Discrete Distributions, Montreal*, 373-378.

Collins, P. y Wyatt, S. (1998), "Citation in Patent to the Basic Research Literature", *Research Policy*, 17, 65-74.

Cornish, W.R. (1997), "Intellectual Property", *Technology and Culture*, 32, 4, 837-845.

Farewell, V.T. (1986), "Mixture Models in Survival Analysis: Are They Worth the Risk?", *Canadian Journal of Statistics*, 14, 257-262.

Gilbert, C.L. (1981), "Econometric Models for Discrete (Integer Valued) Economic Processes". In E.G. Charatsis (ed.) *Proceedings of the Econometric Society European*

Meeting 1979. Selected econometric papers in memory of Stefan Valavanis.
Amsterdam: North Holland.

Gourieroux, C., Monfort, A. y Trognon, A. (1984a), "Pseudo Maximum Likelihood Methods: Theory", *Econometrica*, 52, 681-700.

Gourieroux, C., Monfort, A. y Trognon, A. (1984b), "Pseudo Maximum Likelihood Methods: Applications to Poisson Models", *Econometrica*, 52, 701-720.

Gurmu, S., Rilstone, P. y Stern, S. (1999), "Semiparametric Estimation of Count Regression Models", *Journal of Econometrics*, 88, 123-150.

Hall y otros (1999), "Market Value and Patent Citations: A First Look", *NBER, Working Paper*, June.

Harhoff y otros (1999), "Citation Frequency and The Value of Patented Inventions", *The Review of Economics and Statistics*, 81, 3, 511-515.

Hausman, J.A., Hall, B.H. y Griliches, Z. (1984), "Econometric Models for Count Data with an Application to the Patents-R and D Relationship", *Econometrica*, 52, 909-938.

Heilbron, D.C. (1989), "Generalized Linear Models for Altered Zero Probabilities and Overdispersion in Count Data" unpublished technical report, *Dept. of Epidemiology and Bioestadistics, University of California, San Francisco*.

Joly, P.B. y Looze, M.A. (1996), "An Analysis of Innovation Strategies and Industrial Differentiation through Patent Applications: The Case of Plant Biotechnology", *Research Policy*, 25, 1027-1046.

Lambert, D. (1992), "Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing", *Technometrics*, 34, 1-14.

Lerner, J. (1994), "The Importance of Patent Scope: An Empirical Analysis", *Rand Journal of Economics*, 25, 2, 319-333.

Meyer, M. (2000), "Does Science Push Technology? Patents Citing Scientific Literature", *Research Policy*, 29, 409-434.

Mullahy, J. (1986), "Specification and Testing of Some Modified Count Data Models", *Journal of Econometrics*, 33, 341-365.

Mullahy, J. (1997), "Heterogeneity, Excess Zeros and the Structure of Count Data Models", *Journal of Applied Econometrics*, 12, 337-350.

Narin, F. (1993), "Patent Citation Analysis: The Strategic Application of Technology Indicators", *Patent World*, April, 25-30.

- Narin, F. y Olivastro, D. (1988), "Technology Indicators Based on Patents and Patent Citations", in *Van Raan, A.F.J. (ed.) Handbook of Quantitative Studies of Science and Technology*. North Holland.
- Nelder, J.A. y Wedderburn, R.W.M. (1972), "Generalized Linear Model", *Journal of the Royal Statistics Society A*, 135, 370-384.
- Ozuna, T. y Gómez, I. (1995), "Specification and Testing of Count Data Recreation Demand Functions", *Empirical Economics*, 20, 543-550.
- Patil, G.P. (1970), "Random Counts in Models and Structures", vol 1-3, *University Park, PA, and London, Pennsylvania State University Press*.
- Sánchez Padrón y otros (1999a), "An Analysis of EPO Renewals: A First Step Toward Patent Value Estimation. Some Conceptual Considerations. *Documento de Trabajo n° 98/99-03*. Universidad de La Laguna. (Próxima aparición en McMillan)
- Sánchez Padrón y otros (1999b), "Learning from Patent Survivals: How Serviceable is the Panda Thumb". *Documento de Trabajo n° 98/99-04*. Universidad de La Laguna.
- Schmoch, U. (1993), "Tracing the Knowledge Transfer from Science to Technology as Reflected in Patent Indicators", *Scientometrics*, 26,1,193-211.
- Stauder, D. (1992), "The History of Art. 69(1) EPC and Art. 8(3) Strasbourg Convention on the Extent of Patent Protection", *IIC*, 23, 3, 311-327.
- Trajtenberg, M. (1990), "A Penny for your Cutes: Patent Citations and the Value of Innoventions", *Rand Journal of Economics*, 21,1,172-187.
- Vuong, Q.H. (1986), "Likelihood Ratio Tests for Model Selection and Non-nested Hypothesis", *Econometrica*, vol. 57, 2, 307-333.
- Winkelmann, R. (1995), "Duration Dependence and Dispersion in Count-Data Models", *Journal of Business and Economic Statistics*, 13, 467-474.
- Winkelmann, R. y Zimmermann, K.F. (1991), "A New Approach for Modeling Economic Count Data", *Economics Letters*, 37, 139-143.
- Woolgar, S. (1991), "Beyond the Citation Debate: Towards a Sociology of Measurement Technologies and their use in Science Policy", *Science and Public Policy*, 18,5,319-326.

ANEXO

Tabla 1. *Medidas Descriptivas de las Patentes (N= 629)*

	GR	DS	I1	I2	C
Media	1981.975	6.3339	2.3831	1.7075	2.3847
Desviación Típica	1.5453	1.8877	1.4101	0.8296	3.3707
Mínimo	1980	1	1	1	0
Máximo	1993	10	11	6	23

GR: Fecha de Concesión; **DS:** N° de Estados Designados; **I1:** N° de Clases del IPC (sin restringir); **I2:** N° de Clases a 4 dígitos del IPC; **C:** N° de Citas Recibidas.

Tabla 2. *Distribución de las Citas*

C	Frecuencia	%	% Acumulado
0	196	31.2	31.2
1	149	23.7	54.8
2	92	14.6	69.5
3	57	9.1	78.5
4	35	5.6	84.1
5	27	4.3	88.4
6	17	2.7	91.1
7	9	1.4	92.5
8	12	1.9	94.4
9	6	1.0	95.4
10	8	1.3	96.7
11	3	0.5	97.1
12	3	0.5	97.6
13	1	0.2	97.8
14	3	0.5	98.3
15	2	0.3	98.6
16	2	0.3	98.9
18	3	0.5	99.4
20	1	0.2	99.5
21	1	0.2	99.7
22	1	0.2	99.8
23	1	0.2	100.0
	629	100.0	

Tabla 3. *Estimaciones de los modelos*³⁰

VARIABLES	POISSON	NBI	NBII	ZIP
Constante	-0.3300 (0.148)	0.1344 (0.252)	-0.2854 (0.323)	0.1188 (0.103)
DS	0.0675 (0.015)	0.0244 (0.024)	0.0502 (0.027)	0.0653 (0.009)
I2	0.2251 (0.028)	0.1768 (0.048)	0.2534 (0.068)	0.1814 (0.016)
GR1	0.0813 (0.015)	0.0491 (0.024)	0.0818 (0.034)	0.8618 (0.009)
D2	-0.4254 (0.121)	-0.2762 (0.202)	-0.4104 (0.207)	-0.4104 (0.072)
D3	0.2399 (0.093)	0.2315 (0.155)	0.2431 (0.181)	0.1348 (0.055)
D4	-0.1569 (0.345)	0.5534 (0.601)	-0.1773 (0.934)	-0.1768 (0.359)
D5	-0.4062 (0.212)	-0.1147 (0.372)	-0.3672 (0.425)	-0.4318 (0.182)
D6	-0.4811 (0.155)	-0.2282 (0.255)	-0.4770 (0.306)	-0.4522 (0.136)
D7	-0.1226 (0.128)	-0.4410 (0.217)	-0.0846 (0.243)	-0.1828 (0.081)
D8	-0.1310 (0.132)	0.6721 (0.214)	-0.3396 (0.260)	-0.1291 (0.099)
α	————	2.6933 (0.254)	1.0639 (0.0987)	————
τ	————	————	————	-0.8562 (0.085)
Log-L	-1614.610	-1268.973	-1255.961	-1460.42796
R_p^2	0.2129	-0.2813	0.0432	0.2062

Nota: los valores entre paréntesis se refieren a las desviaciones típicas estimadas

Tabla 4. *Tests de Sobredispersión*

ESTADISTICOS	POISSON/NBI	POISSON/NBII
LR	691.3	717.3
W	10.6	10.8
LM	6.5	7.2

Nota: Bajo H_0 , el estadístico LR se distribuye como una χ^2 ; W y LM como normales estándar

³⁰ También se realizaron las estimaciones de los modelos considerando la variable I2, obteniéndose conclusiones análogas.

Tabla 5. *Frecuencias observadas y ajustadas para los distintos modelos*

Nº CITAS (C)	Nº PATENTES OBSERVADAS	POISSON	BNI	BNII	ZIP
0	196	58	420	192	241
1	149	137	95	129	93
2	92	165	45	90	112
3	57	131	25	63	89
4	35	78	15	45	53
5	27	37	10	32	26
6	17	15	6	22	10
7	9	5	4	16	4
8	12	2	3	11	1
9	6	1	2	8	0
10	8	0	1	6	0
11	3	0	1	4	0
12	3	0	1	3	0
13	1	0	1	2	0
14	3	0	0	2	0
15	2	0	0	1	0
16	2	0	0	1	0
17	0	0	0	1	0
18	3	0	0	0	0
19	0	0	0	0	0
20	1	0	0	0	0
21	1	0	0	0	0
22	1	0	0	0	0
23	1	0	0	0	0
E(C)	2.385	2.396	0.829	2.359	1.625
Var(C)	11.361	2.430	2.950	7.97	2.86
χ^2	—	430.17*	343.9*	12.4	68.25*

Nota: E(C) y Var(C), hacen referencia a las medias y varianzas de las distribuciones de frecuencias observada y ajustadas por los distintos modelos. χ^2 se refiere al test de bondad de ajuste. * indica los modelos en los que se rechaza la hipótesis nula al 5%.